Robot Bullying

Merel Keijsers

Robot Bullying

Candidate	Merel Keijsers Human Interface Technology Lab New Zealand (HIT Lab NZ) College of Engineering University of Canterbury, Christchurch New Zealand merel.keijsers@pg.canterbury.ac.nz merel.keijsers@planet.nl
Primary supervisor	Christoph Bartneck Human Interface Technology Lab New Zealand (HIT Lab NZ) christoph.bartneck@canterbury.ac.nz
Second supervisor	Friederike Eyssel Applied Social Psychology and Gender Research Lab Faculty of Psychology and Sports Science Universität Bielefeld Germany feyssel@cit-ec.uni-bielefeld.de

Start date	April 2017
Submission date	(First version) 23 March 2020 (Final version) 14 July 2020
Oral exam	11 June 2020

There's no difference you can make
And if it seems like an accident
A collage of senselessness
You weren't looking hard enough
I wasn't looking hard enough
at it

Bright Eyes, I Believe in Symmetry

Contents

1 Introduction **9**

1.1 Thesis research questions . . . 9

1.2 General literature review . . . 10

1.2.1 Bullying . . . 11

1.2.2 Robot bullying . . . 12

1.2.3 Perceiving (non)humans . . . 14

1.2.4 Humanness and aggression . . . 17

1.2.5 Current thesis . . . 19

2 Experiment I: Robot bullying versus human bullying **21**

2.1 Introduction . . . 21

2.1.1 Literature . . . 22

2.1.2 Current study . . . 24

2.2 Method . . . 25

2.2.1 Participants and Design . . . 25

2.2.2 Measurements . . . 26

2.2.3 Video material . . . 27

2.2.4 Procedure . . . 28

2.3 Results . . . 28

2.3.1 Preliminary analyses . . . 28

2.3.2 Main analyses . . . 31

2.4 Discussion . . . 33

2.4.1 Limitations . . . 35

2.4.2 Conclusion . . . 36

3 Experiment II: Robot sentience and acceptability of mistreatment **37**

3.1 Introduction . . . 37

3.1.1 Literature . . . 37

3.1.2 Current studies . . . 40

3.2 Experiment II.A . . . 41

3.2.1 Method . . . 41

3.2.2 Results . . . 44

3.2.3 Discussion . . . 46

3.3 Pilots . . . 47

3.3.1 Pilot 1: Robot mind attribution manipulation . . . 47
3.3.2 Pilot 2: Robot bullying . . . 49
3.3.3 Conclusion . . . 50
3.4 Experiment II.B . . . 50
3.4.1 Method . . . 50
3.4.2 Results . . . 53
3.4.3 Discussion . . . 54
3.5 Main discussion . . . 54
3.5.1 Limitations . . . 56
3.5.2 Conclusion . . . 57

4 Study III: The Cleverbot Studies 58
4.1 Introduction . . . 58
4.1.1 Literature . . . 59
4.1.2 Current study . . . 62
4.2 Method . . . 64
4.2.1 Procedure . . . 64
4.2.2 Dataset . . . 64
4.2.3 Coding procedure . . . 65
4.2.4 Variables . . . 65
4.3 Results . . . 67
4.3.1 Preliminary analyses . . . 67
4.3.2 Main analyses . . . 69
4.4 Discussion . . . 71
4.4.1 Limitations . . . 74
4.4.2 Conclusion . . . 75

5 Experiment IV: Mindless robots get bullied 76
5.1 Introduction . . . 76
5.1.1 Current studies . . . 77
5.2 Experiment 1a . . . 79
5.2.1 Method . . . 79
5.2.2 Results . . . 81
5.2.3 Discussion . . . 83
5.3 Experiment 1b . . . 85
5.3.1 Method . . . 85
5.3.2 Results . . . 86
5.4 Discussion . . . 89
5.4.1 Limitations . . . 91
5.4.2 Conclusion . . . 91

6 Experiment V: Teaching robots a lesson **93**
6.1 Introduction . . . 93
6.1.1 Current study . . . 94
6.2 Method . . . 95
6.2.1 Participants and Design . . . 95
6.2.2 Experimental manipulations . . . 96
6.2.3 Procedure . . . 97
6.2.4 Materials . . . 99
6.2.5 Measurements . . . 101
6.3 Results . . . 102
6.3.1 Homogeneity of variance . . . 102
6.3.2 Preliminary analyses . . . 102
6.3.3 Main analyses . . . 103
6.3.4 Model interpretation . . . 105
6.4 Discussion . . . 105
6.4.1 Predictions and findings . . . 106
6.4.2 Strengths and limitations . . . 108

7 Conclusions **110**
7.1 Thesis research questions . . . 111
7.1.1 Is robot bullying seen as fundamentally different from human bullying? 111
7.1.2 Is moral acceptability of robot bullying dependent on mind attribution? . . . 112
7.1.3 Is mind attribution to a robot related to bullying? . . . 113
7.1.4 Does this relationship hold in different contexts? . . . 114
7.2 Problems encountered with the concept "robot bullying" . . . 115
7.2.1 Can robots be bullied? . . . 115
7.2.2 Methodological problems . . . 118
7.3 Empirical issues with anthropomorphism . . . 120
7.4 Mind attribution findings . . . 123
7.4.1 Factors of mind attribution . . . 123
7.4.2 Mind attribution factors across experiments and different robot embodiments . . . 124
7.5 Generalisations and implications . . . 127
7.6 Future research . . . 128
7.7 Last words . . . 129

References **130**

Appendix A Questionnaires **146**

Appendix B Other publications **150**

Appendix C Examiner feedback **152**
C.1 Dr Alan Wagner 152
C.2 Dr Bilge Mutlu 157

Summary

When robots made their first unsupervised entrance to the public space, their engineers were confronted with an unexpected phenomenon: robot bullying (see for example Brscić, Kidokoro, Suehiro, & Kanda, 2015; Salvini et al., 2010). While the phenomenon has continued to manifest itself since and a few theoretical explanations have been suggested, little empirical work has been done to substantiate any theorising as of yet.

This thesis summarizes five pieces of research that explore what psychological factors influence people's willingness to behave anti-socially towards robots. It is structured around four experiments on human-robot interaction (Chapters 2, 3, 5, and 6) and one analysis of human-chatbot interaction (Chapter 4). In addition, there are some general reflections on the methodological and philosophical issues with studying robot bullying (section 7.2), as well as the role of mind attribution (i.e., attributing the ability to think and feel to another being; section 7.4), which has been a recurring measure of interest throughout the experiments.

Chapter 1 provides an overview of the motivation for the thesis topic and the research questions. It also includes a general discussion of the relevant literature, focusing on anthropomorphism of nonhuman agents, mind attribution as a factor of anthropomorphism, and how dehumanisation as a facilitator for interhuman aggression may be generalisable to human-robot interaction as well.

Chapter 2 describes an experiment that explored whether bullying behaviour is perceived as more morally acceptable if the victim is a robot rather than a human. The results indicated no significant difference in moral acceptability, and suggested that higher levels of mind attribution were related to lower acceptability of abuse.

Chapter 3 expands on these findings by describing two studies that experimentally manipulated mind attribution. Also, whereas participants in the experiment from Chapter 2 were passive spectators of a human-robot interaction, one of the experiments in this chapter involved active interaction between a participant and a robot. In two experiments we investigated the influence of a robot's mind attribution on the perceived acceptability of robot bullying and people's willingness to bully a robot. Results indicated that acceptability of robot bullying can be manipulated both explicitly, by providing people with information on the robot's mind attribution, and implicitly, through having the robot give off emotional cues. Those effects are independent of one another. Interestingly, robot mind attribution was not associated with a lower robot bullying incidence rate in this experiment.

In contrast to the studies reported in the other chapters, the study covered in Chapter

4 did not realise an experimental design. Almost 300 conversations between users and an online chatbot were harvested and coded for humanlikeness of the chatbot, self-disclosure by the user, and importantly, the amount of verbal abuse or sexual harassment. Subsequent analyses showed that humanlikeness in the chatbot was associated with more abuse (both sexual harassment and verbal aggression). Self-disclosure in terms of making mention of one's gender (both male and female) was associated with less verbal aggression, but more sexual harassment.

Chapter 5 describes an experiment which investigated whether mind attribution is linked to robot abuse. Mind attribution to the robot was intended to be manipulated through priming participants with a feeling of power, as previous studies on dehumanisation had shown that power reduces mind attribution. In addition, humanlike qualities of the robot were manipulated. The participants' verbal abuse of a virtual robot was measured as the main outcome of interest; mind attribution to the robot and humanlikeness of the robot were measured as manipulation checks. Contrary to previous findings in human-human interaction, priming participants with power did not result in reduced mind attribution. However, evidence for dehumanisation was still found, as the less mind participants attributed to the robot, the more aggressive responses they gave. This effect was moderated by the power prime and robot humanlikeness manipulation.

The discussion section of Chapter 5 offers an explanation for the surprising results, which is put to the test in Chapter 6, where an expansion of the experiment from Chapter 5 is presented. Feelings of power, robot embodiment (virtual versus embodied) and feelings of threat were experimentally manipulated. Participants played a learning task with either a virtual or an embodied robot, and were asked to restrict the robot's energy supply after each wrong answer, which was taken as a measure of aggression. Results indicated that an embodied robot was punished less harshly than a virtual one, except for when people had been primed with power and threat. Being primed with power diminished the influence of mind attribution on aggression. Mind attribution increased aggression in the threat condition, but was related to decreased aggression when people had not been reminded of threat. These results suggest that while mind attribution appears to play a role in robot bullying, the relationship is too complicated to be explained by dehumanisation theory alone.

Finally, Chapter 7 aggregates the results from the studies in this thesis to provide an answer to the thesis research questions. In addition, the strengths and limitations of the research are discussed. Furthermore, trends in mind attribution to the robots used in the different experiments are discussed. Finally, possible directions for future research are considered.

Chapter 1

Introduction

Over a decade ago, a small cleaning robot was assigned an unsupervised job at a public square while scientists were conducting a field experiment with a larger and more sophisticated robot nearby. However, things did not go quite as planned. The cleaning robot was approached by random bystanders and abused — in the absence of any form of provocation from the robot. The baffled scientists had to hastily reach for their cell phones to capture this unexpected form of human-robot interaction (Figure 1.1, leftmost image). Salvini et al. (2010) noted that *"the nature of the abuses suffered by the robots [...] is much more similar to bullying behaviours than vandalism. [...] In the case of urban robots, acts of vandalism could be, for instance, crashing the touch screen monitor, setting fire to the robot, or keying the robot cover. On the contrary, what we noticed during the behavioural study were actions aimed at forcing the robot to do or not do something or, in a few cases, simulations of "physical" attacks"* (Salvini et al., 2010, p. 371).

A few years later, in 2014, researchers struggled to keep a mall robot safe from being attacked by children. In spite of the robot tending to its own business and not interacting with the mall visitors, children would flock around it and kick it, push it, and insult it (see Figure 1.1, centre image). The engineers working on the project tried different strategies to stop the children from attacking the robot, but to no avail. Eventually, they had to resort to conflict avoidance: whenever the robot detected a human that was too short to be an adult, it turned and ran (Brscić et al., 2015).

Even more recently, in 2017, an intoxicated man was arrested for assault in a car park in Mountain View, California. A local commented on the incident, stating "*I think this is pretty pathetic (...) because it shows how spineless drunk guys (...) really are because they attack a victim who doesn't even have any arms. I don't think this is a fair fight, really totally unacceptable.*" Fortunately, the victim – a K5 Knightscope robot (see Figure 1.1, rightmost image) — only suffered minor scratches (Vincent, 2017) and was back on patrol a few days later.

1.1 Thesis research questions

The cases described above provided the topic for the current thesis on robot bullying. The main research question is to what extent mind attribution (i.e., perceiving the robot

Figure 1.1: From left to right: Adults kicking a cleaning robot (Salvini et al., 2010); Children ambushing a shopping mall robot (Brscić et al., 2015); The K5 Knightscope robot (source: Lasica, 2014).

as being capable of thinking and feeling (Gray, Gray, & Wegner, 2007)) influences robot bullying tendencies. This broad question was further refined into the following research questions:

1. Is robot bullying seen as fundamentally different from human bullying?
2. Is the moral acceptability of robot bullying dependent on mind attribution?
3. Is mind attribution to a robot related to robot bullying?
4. Does this relationship hold in different contexts?

1.2 General literature review

Humans recognise robots as social agents. They talk to them (Bartneck, Van Der Hoek, Mubin, & Al Mahmud, 2007) as if they understand what is being said, they punish them when they prove to be a bad teammate (Bartneck, Reichenbach, & Carpenter, 2008) — but also feel sorry for them when they are being punished (Slater et al., 2006). They even try to prevent robots from getting hurt (Slater et al., 2006), even though they rationally acknowledge that the robot in question would be incapable of feeling and does not possess awareness (Darling, 2012). This tendency to see robots as social agents also shows in cognitive responses; for example, humans interpret a robot's behaviour along (human) stereotypes (Bartneck et al., 2018; Eyssel & Hegel, 2012). Our brain responds to robots as if they were giving off social cues; activating mirror neurons when watching a robot perform an action (Gazzola, Rizzolatti, Wicker, & Keysers, 2007), activating neural networks linked to the theory of mind when playing a game with a robot (Krach et al., 2008), and activating areas associated with emotional empathy when watching a robot getting hurt (Rosenthal-von der Pütten, Krämer, Hoffmann, Sobieraj, & Eimler, 2013). At the physiological level, participants' heart rates and skin conductance level increased when they had to administer increasingly heavy shocks to a virtual agent in an adaptation of Milgram's obedience studies (Slater et al., 2006), indicating increased arousal when participants had to "punish" the agent. The display of distress was echoed in participants' self-reported stress levels,

with participants reporting being more self-aware and having higher levels of physiological stress indicators such as trembling and feeling hot in the face after the "learning task" had finished compared to pre-task measurements. In short, humans respond with social cognition, social affect, and social behaviour when interacting with robots.

However, not all social behaviours are positive.The Knightscope (Vincent, 2017) and the cleaning robot (Salvini et al., 2010) are not alone in being attacked by humans (see for example Brscić et al., 2015; Mutlu & Forlizzi, 2008; Nomura, Kanda, Kidokoro, Suehiro, & Yamada, 2016; Rehm & Krogsager, 2013). Abusive behaviour towards robots ranges from more playful forms of mocking and name-calling (see for example Rehm & Krogsager, 2013) to sabotaging the robot's goals and obstructing its way (Brscić et al., 2015; Nomura et al., 2016) to physical assault like kicking and slapping (e.g. Mutlu & Forlizzi, 2008; Salvini et al., 2010), and has been reported across cultures and age groups. As Salvini et al. (2010) remarked, this behaviour appeared to be motivated by the wish to engage with the robot in a social way (albeit negative) rather than representing an act of vandalism. Vandalising robots would have as main objective to damage the robot, and one would thus expect people to set fire to them, key them, or attempt to crash their interface. Instead, humans assault robots in a similar way as they bully sentient creatures - by kicking and insulting them or trying to force or obstruct their actions (Brscić et al., 2015; Salvini et al., 2010). As a consequence, this negative behaviour towards robots has been labelled *robot bullying*, a term later adopted by other HRI researchers (see for example Ku, Choi, Lee, Jang, & Do, 2018; Nomura et al., 2016; Tan, Vázquez, Carter, Morales, & Steinfeld, 2018).

1.2.1 Bullying

In spite of being a widely adopted word in both academic and everyday language, the term "bullying" is surprisingly ill-defined. There doesn't appear to be consensus on what exactly bullying is (and isn't) (Hamburger, Basile, & Vivolo, 2011; Modecki, Minchin, Harbaugh, Guerra, & Runions, 2014). Some authors avoid the problem of finding a generally accepted definition altogether by never bothering to define what they understand as bullying in their research on the topic (Cowie & Berdondini, 2002; Fox & Boulton, 2005; Ireland & Monaghan, 2006; Mishna, Schwan, Lefebvre, Bhole, & Johnston, 2014, for example). This rather complicates deciding on a comprehensive definition.

In spite of the variety in definitions of bullying, there are a few characteristics that emerge in the majority of them. Most scholars (see for example Ang & Goh, 2010; Casper, Meter, & Card, 2015; Gullone & Robertson, 2008; Hamburger et al., 2011; Jolliffe & Farrington, 2011; Modecki et al., 2014; Postigo, González, Montoya, & Ordoñez, 2013; Sokol, Bussey, & Rapee, 2016) include the following components:

- physical and/or psychological aggression that's intended and repeated over time;
- and which occurs in a dominant/submissive relationship, with a power imbalance between the bully (dominant) and the victim (submissive);
- and has the goal of harming or hurting the victim.

Some authors have added to this list that the bullying behaviour must be unprovoked by the victim (Gullone & Robertson, 2008; Postigo et al., 2013); that the behaviour must be proactive, with the bully going out of their way to seek out the situation (Jolliffe & Farrington, 2011); or that the bully must enjoy the bullying (Gullone & Robertson, 2008). Others have claimed that repetition and power imbalance, with the bully being in a position of power over the victim, are in fact not key defining features (Lowry, Zhang, Wang, & Siponen, 2016; Modecki et al., 2014; Volk, Veenstra, & Espelage, 2017).

What's furthermore problematic is that both the power imbalance and the repetition are somewhat open to interpretation (Casper et al., 2015). The power imbalance may be pre-existing, but it can also be considered the goal of the bully (Volk et al., 2017). The initial act of aggression could be used by the bully to test whether a peer is susceptible to intimidation; if this turns out to be the case, the power dynamic is established through the perpetrators repeated aggression (Casper et al., 2015). Of course, robots are the ideal target for bullying as they are in a clear subordinate position, will not retort in kind, and cannot feel any pain, which absolves the aggressor from any moral consequence (De Angeli, Brahnam, Wallis, & Dix, 2006). Thus, some power imbalance is already in place in HRI, and bullying behaviour from the human may capitalize on and enhance this pre-existing inequality.

In addition, "repeated over time" has been operationalised in measurements as: more than once in a 7-day (Bond, Wolfe, Tollit, Butler, & Patton, 2007), 30-day (Bosworth, Espelage, & Simon, 1999), or even 6-month (Sokol et al., 2016) time window; while other instruments simply report on a scale ranging from "never" to "all of the time" (Crick & Grotpeter, 1995). Thus, the repetition itself rather than the time window in which this occurs appears to be critical in defining bullying. The repetitive element is taken as an operationalisation that ensures that the intention of the aggressor was indeed malicious (rather than, say, an ill-received joke, a one-off outburst of frustration, or even curiously "poking the bear" to see what the response would be).

For all practical purposes, in this thesis bullying is defined as consisting of the following three components: physical and/or psychological aggression, which occurs in a dominant/submissive relationship, and has the goal of harming or hurting the victim. All but one (Experiment II, Chapter 3) of the experiments reported in this thesis incorporate a repetitive aspect of bullying. Thus, it is possible to differentiate between one-off instances of negative behaviour, which may be performed out of curiosity rather than malice, and persistent bullying. In addition, in all experiments a non-aggressive alternative for the aggressive behaviour was present. This means that participants deliberately chose bullying behaviour over a more agreeable alternative, thus further ensuring the component of intention. See section 7.2 for a more in-depth discussion of the problems encountered with the concept "robot bullying".

1.2.2 Robot bullying

Considering how robots are (as of yet) incapable of getting hurt by any bullying behaviour, one could wonder why robot bullying would be considered problematic. However, from

an ethical perspective, some behaviours can be deemed immoral even if performed on an entity that is incapable of any suffering, like a robot (Sparrow, 2017). Since the robot is recognised by the human as a social actor, abusing it might encourage treating other humanlike beings (e.g. actual humans) in a similar way (Whitby, 2008). More generally speaking, the assertion "I can do whatever I desire with a robot" rests upon the idea that all and any actions are acceptable as long as no-one gets harmed (Richardson, 2016), which even in the most libertarian societies is not a commonly shared attitude (Whitby, 2008). For example, one could argue that it is okay to privately kick and hit a corpse since it cannot feel, as long as the body does not get visibly damaged. Yet this is a highly provocative statement.

But robot abuse should be considered problematic from a pragmatic point of view as well. Both scholars and the industry believe that within a few decades robots will play a major role in society (Cooper, 2019; Walsh, 2018). Understanding the determinants of robot abuse will help to develop strategies to prevent, discourage, and respond to robot bullying. These strategies will be needed because robot abuse may lead to a malfunctioning robot which will likely be expensive to replace or to repair, and might create hazardous situations for the abuser, bystanders, and future users (De Angeli et al., 2006).

A few studies have been published on the development of strategies to prevent and discourage robot bullying. But in spite of researchers' best efforts to design behaviours which discourage robot bullying, it has been shown to be remarkably persistent (see for example Bartneck & Hu, 2008; Brscić et al., 2015; Nomura et al., 2016; Salvini et al., 2010). Prevention strategies to reduce robot-directed aggression include making the design of the robot so robust that it simply cannot be damaged (Salvini et al., 2010); having the robot shut down completely for either a specified period of time (Ku et al., 2018) or until the abusive behaviour stops (Tan et al., 2018); or having the robot run away from humans that are under 1.40m tall, since children are more likely than adults to engage in abuse (Brscić et al., 2015). Some of these strategies have shown to be moderately successful in reducing robot abuse (Brscić et al., 2015; Ku et al., 2018). However, they do not target the underlying reason for the bullying. In addition, none of them is particularly useful in a situation where robots become ubiquitous, such as autonomous driving cars or robots patrolling public spaces. In order to develop more effective strategies, a deeper understanding of the motivation behind robot bullying is needed.

Research on the reasons behind robot bullying is still sparse (De Angeli & Brahnam, 2008) and often involves anecdotal observations (e.g. Brscić et al., 2015; Mutlu & Forlizzi, 2008; Rehm & Krogsager, 2013; Salvini et al., 2010). Initial studies pitched an evolutionary explanation (De Angeli & Brahnam, 2008), suggesting that when you come across a creature you have never encountered before, poking and prodding it is one way to figure out how responsive and potentially dangerous it is. Robot bullying then would be humans testing the robot's boundaries in order to learn how best to behave around it. Others suggest that disinhibition for aggressive behaviour to occur when "the illusion of anthropomorphism shatters" and the human suddenly stops seeing the robot as a social agent (Bartneck, Rosalia, Menges, & Deckers, 2005; De Angeli, 2006). This explanation

suggests that humans initially expect the robot to behave in a humanlike way. When the robot inevitably fails to meet this expectation, humans realise they no longer have to be polite to the robot and thus get abusive. Bartneck et al. (2008) hypothesised that abuse might be caused by frustration, if a robot does not respond as expected.

Further research is needed to shed light on the determinants of robot bullying, particularly to enable effective interventions (see also Brahnam & De Angeli, 2008; Eyssel, 2017). The current thesis will thus dive into the psychological motivations behind robot bullying behaviour. The goal is to experimentally test whether aggressive behaviour towards robots is a social phenomenon, and is guided by the same social processes as aggressive behaviour towards humans.

1.2.3 Perceiving (non)humans

In 1994, Nass, Steuer, and Tauber (1994) found that humans treat computers as if they are social actors. In a following series of experiments, Reeves and Nass (1996) replicated a slew of well-established social mechanisms from human-human interaction in interactions where one of the interaction partners was substituted with a computer. For example, they showed that humans will be more polite in their feedback of a computer if they have to input said feedback on the same computer as they are reviewing; that humans consider it rude for a computer to sing its own praise but are accepting of a different computer complimenting the first computer on a job well done; and that, following human stereotypes, a "male computer" is more convincing in its praise than a "female computer". These findings inspired the Media Equation theory, which states that humans will automatically respond to media as if it is real life (Reeves & Nass, 1996). Later studies further confirmed that people interact with machines and media as if they are social agents (Luczak, Roetting, & Schmidt, 2003).

The Media Equation has been extended to robots as well. When playing a cooperative game with either a robotic or a human partner, participants apply the same social norms to both partners, punishing bad performance and rewarding good performance (Bartneck et al., 2008). Eyssel and Hegel (2012) found that a "male" robot (with a short haircut) was rated as higher in agency, while a "female" robot (which looked and acted exactly the same as the male version, except for having long hair) as being warmer and more social; in addition, the short-haired "male" robot was considered more suitable for stereotypically male tasks such as transporting goods and monitoring technological devices, whereas the long-haired "female" robot was seen as more capable of stereotypically female tasks such as elderly care and meal preparation (Eyssel & Hegel, 2012). The Media Equation in robots can be seen in the brain's response to robots as well. Seeing a robot hand carrying out a goal-oriented series of movements (e.g. picking up an object) activates the same mirror neurons in the brain as observing a human hand performing the same action (Gazzola et al., 2007; Oberman, McCleery, Ramachandran, & Pineda, 2007). Moreover, Krach et al. (2008) compared brain activation when participants were playing a competitive game with either a computer, a functional robot (built out of LEGO), a robot with a humanlike build including a head and face, or a human opponent. With increasing humanlikeness of the

opponent, activation of brain regions that are associated with the theory-of-mind neural network became enhanced (Krach et al., 2008). At least partially, the media equation thus seems to be an expression of anthropomorphism.

Anthropomorphism (from the Greek words *anthropos*, meaning "human", and *morphe*, meaning "form" (Lesher, 2001)) is the phenomenon of attributing human characteristics, emotions, and motivations to nonhuman agents (Epley, Waytz, & Cacioppo, 2007). These nonhuman agents can be concrete and tangible, like animals or robots, or abstract and philosophical, such as the wind, sea, spirits and deities. Moreover, "attributing human characteristics, emotions, and motivations" may refer to looking or sounding humanlike (as is the case with android robots or apes) as well as being attributed humanlike mental capacities (such as having emotions, awareness, and plans). As of such the term is quite broad and two people could use the term correctly while talking about wildly different situations (Fisher, 1995, see also section 7.3 for a discussion of the implications).

In the light of this potentially confusing broadness, it is useful to narrow down what will be meant with "anthropomorphism" in the rest of this thesis. It has been proposed that there are two broad types of anthropomorphism: interpretative and imaginative anthropomorphism (Fisher, 1995). Although this framework is not commonly used in the field of HRI, it actually can be quite helpful to further specify whether anthropomorphism as a term is referring to mental capacities, humanlike appearance, or both. Interpretative anthropomorphism concerns the inference of mental states to an agent by interpreting its behaviour in a similar way as we would interpret human behaviour. For example, when the dog comes running at us when we arrive back home after a day out, we may anthropomorphise the behaviour by saying *"ahhh look, he missed us"*. Imaginative anthropomorphism, on the other hand, is using human characteristics as a template when imagining agents that have never been encountered by anyone. For example, depicting God as an old white man with a beard who communicates through spoken language, or aliens like grey bipeds with oversized heads whose sole mission is to colonise and violently subdue the natives in any new territory they happen to stumble upon, would be instances of (white) people anthropomorphising the concept of "deity" and "alien". In the current thesis, only interpretative anthropomorphism will be considered. Thus, when the term "anthropomorphism" is used, it will refer exclusively to a person inferring mental states to a nonhuman.

To explain the wide range of nonhuman agents that get anthropomorphised, Epley et al. (2007) drew a psychological framework for anthropomorphism that rests on three factors. The first factor considers the cognitive response to the agent. Knowledge about humans, and especially knowledge about the self, can be used as a starting point to make inferences about properties of others. If the agent resembles humans in appearance and behaviour, it is more likely that knowledge on humans will automatically become activated in the brain and used as a base for judging the behaviour of that agent. The other two factors are motivational mechanisms. The first of those is effectance, or the need to understand and predict the behaviour of other agents. Seeing nonhuman agents as possessing human needs and desires can help explain why they behave in the way they do, and create a

sense of confidence when predicting the agent's responses to future events. The second motivational mechanism is sociality, or the need to be social and have relationships with others. By perceiving nonhuman agents as humanlike, people can fulfil the need to have a connection to that agent.

The validity of the agent factor was experimentally confirmed by Eyssel, Kuchenbrandt, Bobinger, de Ruiter, and Hegel (2012), who showed that the relationship between psychological closeness to and the anthropomorphism of robots depends on the robot's voice. When the robot had a synthesised voice, the gender of the robot's voice didn't influence anthropomorphism ratings in participants; but when it had a human voice, robots with a voice that matched the participants' gender were rated as more anthropomorphic. Gender is a fundamental social category: children self-categorise based on gender before any other characteristic (David, Grace, & Ryan, 2004). Perceiving the robot as "being of one's own gender" thus enhanced psychological closeness, but only if the robot sounded humanlike. The findings from Eyssel, Kuchenbrandt, Bobinger, et al. (2012) indicate the importance of the human's attributes as a moderator of psychological closeness to the agent.

Waytz, Morewedge, et al. (2010) confirmed the validity of the second factor in the human-robot interaction field by showing that robots behaving in an unpredictable way have higher anthropomorphism ratings (see also Eyssel & Kuchenbrandt, 2011; Eyssel, Kuchenbrandt, & Bobinger, 2011). Moreover, when participants are dealing with unpredictable robots, a brain region that is involved with inferring mental states of other agents becomes activated, suggesting that the robots are indeed recognised as having a mind of their own (Gazzola et al., 2007).

The third factor was confirmed when loneliness was shown to correlate with the tendency to assign a mind to interactive gadgets (like an alarm clock that "runs away" when it goes off). Moreover, after a feeling of loneliness was experimentally induced, participants had a greater tendency to anthropomorphise a wide range of nonhuman agents: pets, God, and a series of ambiguous drawings in which one might perceive the features of a face (Epley, Akalis, Waytz, & Cacioppo, 2008). Eyssel and Reich (2013); Reich and Eyssel (2013) replicated these finding specifically with social robots.

The Media Equation thus appears to be the consequence of people anthropomorphising computers and machines. However, anthropomorphism as it occurs in the Media Equation still has decidedly different behavioural consequences compared to anthropomorphism as it occurs in animals. While humans apply certain social norms when interacting with a robot, they also display certain behaviours that would be unacceptable in human-anthropomorphic animal behaviour (Bartneck, Van Der Hoek, et al., 2007). For example, no one would be surprised when someone switches off a robot after they are done using it; whereas most people would object to killing a dog or even "just" putting it to sleep by tranquilizing it any time they don't want to play with it (Bartneck & Hu, 2008).

A related theoretical framework called mind perception provides a possible explanation for how animals and robots are anthropomorphised differently. While mind perception is at the core of anthropomorphism, perceiving an agent as capable of cognition and affect

is a dimension of anthropomorphism. It does not encompass anthropomorphism entirely, however (Epley et al., 2007), as it for example does not account for imaginative anthropomorphism. Mind attribution refers to the psychological response where someone makes an inference about the agent's capability to think and feel (Eyssel, 2017). Attributing mind to an agent thus is a specific expression of anthropomorphising it.

The perception of a mind in other agents, human or not, has been shown to consist of two dimensions: Experience and Agency (Gray et al., 2007). The Experience dimension entails to what extent an agent is thought to be capable of experiencing thoughts, feelings, and the world around it. Agents who are capable of Experience and but are not too capable of Agency are perceived as being less responsible for their actions, more prone to feeling hurt, and thus deserving of protection. The Agency dimension indicates to what extent the agent is seen as capable of self-control, memory, planning, and moral judgement. Beings highly capable of Agency but relatively incapable to Experience are considered to be responsible for their own actions and less deserving of protection from harm, as they don't have (a rich set of) feelings and supposedly can take care of themselves. These two dimensions are not mutually exclusive: humans are seen as being highly capable of both Agency and Experience (Gray et al., 2007).

In the survey of Gray et al. (2007), robots were seen as highly capable of Agency but not of Experience, whereas animals were perceived as very capable to Experience but not of Agency (this study however stems from 2007. Robots nowadays may be seen as more capable to Experience). This would explain why robots and animals are not anthropomorphous in the same way: since robots are seen as high on Agency instead of Experience, one can harm them without feeling bad (Gray et al., 2007). Indeed, increasing Agency traits in a robot did not do much for its anthropomorphism or likeability ratings, but increasing its Experience traits resulted in it being perceived as more anthropomorphic and likeable (Salem, Eyssel, Rohlfing, Kopp, & Joublin, 2013; Złotowski, Strasser, & Bartneck, 2014). How anthropomorphic a robot is perceived to be influences the expectations we have of the robot (Darling, 2015; Malle, Scheutz, Forlizzi, & Voiklis, 2016) and what behaviour we deem acceptable towards it (Lucas, Poston, Yocum, Carlson, & Feil-Seifer, 2016; Tan et al., 2018).

So, in summary: the Media Equation states that humans will respond to robots as if they are social agents. This behaviour is a consequence of humans anthropomorphising robots, in terms of attributing humanlike capabilities to them. Previous research suggests this is mostly the capability to think and reason. However, the perceived capability of a robot to fully experience the world around it can be tweaked as well. This suggests that in order to explain robot bullying behaviour, one should look at which psychological mechanisms come into play when humans set out to hurt each other.

1.2.4 Humanness and aggression

People strive to have a positive view of themselves (Shrauger, 1975), perceiving themselves as a virtuous, capable, and pleasant person to be around (Dunning, 1999). Purposefully inflicting hurt or pain upon another sentient being clashes with this self-image, creating a

tension that can be relieved through either adjusting the hurtful behaviour, or by changing the way the behaviour is interpreted (Bem, 1972). Dehumanisation theory explains the cognitive mechanisms that people can use to re-interpret hurtful behaviour so that they maintain a positive self-image. It describes the psychological process by which humans perceive their victims as slightly less capable of thinking and feeling, which decreases the moral standing of the victim and allows the perpetrator to disregard the negative consequences of their own behaviour (Castano & Kofta, 2009; M. Y. Li, Leidner, & Castano, 2014). As a result, the threshold for inflicting pain (both physical and mental) on others is lowered (Haslam, 2006; Haslam & Loughnan, 2014). Research on the role of dehumanisation in human-human interaction has confirmed that a reduced perceived capability to think and feel (Haslam & Loughnan, 2014; Kozak, Marsh, & Wegner, 2006) is related to an increase in aggression (Haslam & Loughnan, 2014; Leidner, Castano, & Ginges, 2013). However, it should be noted that while dehumanisation theory can be applied to bullying behaviour and other forms of aggression, it also covers less ominous instances of humans hurting another. The most common example is doctors dehumanising their patients so they can administer painful (yet effective) treatments (Schulman-Green, 2003).

Haslam (2006) distinguishes between two dimensions of humanness: Human Nature and Uniquely Human traits (see also Haslam, Loughnan, Kashima, & Bain, 2008). Uniquely Human traits and capabilities are presumably reserved for humans only, like higher forms of cognition. Human Nature on the other hand involves traits and capabilities that are shared with other animals, but are at the same time considered fundamental to being human, like fear or joy. Perceiving fewer Human Nature traits in an agent results in a "mechanistic" form of dehumanisation; in humans, this form of dehumanisation is applied in for example the stereotypical banker or businessmen. Alternatively, perceiving less Uniquely Human traits results in an "animalistic" form of dehumanisation (Haslam, 2006; Haslam, Loughnan, et al., 2008); this form of dehumanisation is commonly applied to women or the mentally disabled. Interestingly, although agents high on Human Nature traits are seen as more deserving of protection, animalistic dehumanisation is still related to a decrease in empathy (Castano, Kofta, Čehajić, Brown, & González, 2009) and both types of dehumanisation are related to increased aggression (Haslam & Loughnan, 2014; Leidner et al., 2013).

Dehumanisation can be triggered by stable factors like trait characteristics of the person who dehumanises (e.g. narcissism, extraversion; Kteily, Bruneau, Waytz, & Cotterill, 2015; Locke, 2009) and of the victim (e.g. gender, social class; Bain, Park, Kwok, & Haslam, 2009; Rudman & Mescher, 2012). But circumstantial factors such as a feeling of social connection (Waytz & Epley, 2012), a sense of power (Gwinn, Judd, & Park, 2013; Lammers & Stapel, 2011), or self-focus (Haslam & Loughnan, 2014) can also increase dehumanisation tendencies.

Dehumanisation theory shows considerable overlap with anthropomorphism, to the point where it has been suggested that they are two approaches of the same concept (Waytz, Epley, & Cacioppo, 2010). Loughnan et al. (2010) empirically applied a dehumanisation framework on non-human agents and showed that Uniquely Human traits were

more readily associated with robots, while Human Nature traits were more easily linked to animals. However, not all scholars agree that dehumanisation and anthropomorphism are each other's reverse. For example, in one study neither robot appearance nor perceived intentionality influenced the mind or moral agency attributed to it. A more human appearance of the robot resulted in an increase in ascribed Uniquely Human and Human Nature traits. Yet at the same time, perceived intentionality correlated with a *lower* attribution of Uniquely Human traits (Złotowski, Sumioka, Bartneck, Nishio, & Ishiguro, 2017). These results are surprising because one would expect that if a humanlike appearance enhanced the ascription of Uniquely Human and Human Nature traits, it should have increased mind attribution in a similar fashion; and any relation between intentionality and Uniquely Human traits should have been positive.

The inconsistency in findings might be explained by the many different approaches that have been used to measure dehumanisation and anthropomorphism. Anthropomorphism has been operationalised in a wide variety ways by different researchers (Kätsyri, Förger, Mäkäräinen, & Takala, 2015; Ruijten, Bouten, Rouschop, Ham, & Midden, 2014). There is the humanness subscale of the revised Godspeed questionnaire (Bartneck, Kulić, Croft, & Zoghbi, 2009; C.-C. Ho & MacDorman, 2010), which quite straightforwardly asks the participant to rate the robots on scales like 'living versus inanimate' and 'human-made versus humanlike' (MacDorman & Chattopadhyay, 2016; Złotowski et al., 2014). Other researchers decided to craft their own questionnaires (Eyssel & Pfundmair, 2015; Rosenthal-von der Pütten et al., 2013). Mind attribution questionnaires as well have been used to measure anthropomorphism (Epley et al., 2008; Eyssel, Kuchenbrandt, Bobinger, et al., 2012; Waytz, Morewedge, et al., 2010), as well as Uniquely Humans and Human Nature trait attribution measurements (Loughnan & Haslam, 2007; Salem et al., 2013), which hold the (implicit) assumption of dehumanisation and anthropomorphism being each other's opposite. This makes it virtually impossible to tease apart anthropomorphism, dehumanisation, and mind attribution in the literature. See section 7.3 for a more in-depth analysis of this problem.

1.2.5 Current thesis

The objective of this thesis is by no means to add to or even settle this debate. However, in the light of the ongoing discussion it is important that mind attribution, anthropomorphism, and dehumanisation are not used interchangeably, even though they appear to overlap. For the sake of consistency, most research discussed in this thesis is concerned with mind attribution, which in turn was operationalised as 'the perceived capability of cognition and emotion'. However, based on dehumanisation theory and the mind attribution framework, hypotheses can be drawn for the thesis research questions as proposed in section 1.1.

Based on the Media Equation, it is expected that human bullying and robot bullying will be seen as equally unacceptable. In accordance with the mind perception framework by Gray et al. (2007), it is furthermore hypothesised that the attribution of mind to a robot will cause people to see robot bullying as less acceptable. Thirdly, based on dehu-

manisation theory it is expected that mind attribution to robots will be inversely related to bullying. Contexts that influence dehumanisation and mind attribution tendencies would be expected to influence this relationship.

A series of experiments was conducted to empirically test these hypotheses. Those experiments will be covered in the next five chapters.

Chapter 2

Experiment I: Robot bullying versus human bullying

This chapter is an adapted version of the original paper 'The morality of abusing a robot'. Bartneck, C. & Keijsers, M. (2020). *Paladyn, Journal of Behavioural Robotics*. doi: 10.1515/pjbr-2020-0017

2.1 Introduction

The first thesis research question to be addressed is to what extent robot bullying differs from human-on-human bullying. The Media Equation (Nass et al., 1994; Reeves & Nass, 1996) would suggest that people don't perceive robot bullying as fundamentally different from human bullying. However, to our knowledge, no experimental research to date has directly compared differences in perception of human bullying and robot bullying.

Such a comparison is needed to validate Experiments II through V (Chapters 3 to 6), as a major assumption in these studies is that robots are considered social agents by the participants, and that robot bullying is perceived as a social behaviour. Experiment I therefore compared participants' responses to robot bullying with responses to human bullying. Specifically, it tested whether both types of bullying were seen as equally morally (un)acceptable; and if the acceptability was related to mind attribution for both types of victim.

We would like to acknowledge Jake Watson and Sam Gorski from Corridor Digital who made the stimuli for this experiment available.

Figure 2.1: Kicking a human and a robot in the back

2.1.1 Literature

Researchers are facing several problems when trying to investigate abusive behaviour towards robots (see also section 7.2). Most glaringly, people are unlikely to bully robots during a controlled experiment, as they tend to be self-aware and motivated to make a good impression (Nederhof, 1985; Ritter & Eslea, 2005). Participants have been coerced to physically harm a robot (Bartneck, Verbunt, Mubin, & Mahmud, 2007), but this experiment used explicit instructions to destroy the robots and the robots in question were very simple and cheap. Therefore, the behaviour that this experiment measured may have been obedience rather than robot abuse. It is uncertain if the results would generalise to robot bullying, or even obedience behaviour with a more advanced and anthropomorphic robot.

Some HRI studies on robot bullying have adopted less destructive forms of abuse, such as reducing a robot's electrical power supply (Bartneck, Van Der Hoek, et al., 2007; Keijsers, Kazmi, Eyssel, & Bartneck, 2019) or focusing on the use of abusive language (Keijsers & Bartneck, 2018). Using the framework of Game Theory, robots have also been withheld points or money (Sandoval, Brandstetter, & Bartneck, 2016).

An advantage of these milder forms of abuse is that they can also be employed in comparison studies with humans. Withholding money from participants is a methodology that can still pass an ethics board. Kicking and hitting a participant would most certainly not. But these mild forms of aggression pose a problem of their own. Aggression towards humans can be measured through small transgressions of social norms; rude behaviour that won't cause any physical or severe psychological harm, but are enough to slightly sting. Robots, however, are not sentient and most people are rationally well aware of this. Previous research has suggested that this does not prevent people from automatically applying social behaviour to robots (Reeves & Nass, 1996), but one could still argue that a participant omitting any polite conversation or withholding any reward (monetary or otherwise) from a robot is the result of them reasoning that the robot could not care less about whether a command ends with 'please' or if it's awarded any payment. Ultimately, one could wonder if robot abuse would be measured, or the participant's desire to come across as a rational human being.

Alternatively, more extreme abusive behaviours can be studied when participant's responses to recordings of abuse are measured. Common examples of this method are vignette-based approaches, where participants read about abusive behaviour and then express their moral sentiment towards the actions described (e.g. Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015), video-based approaches, where participants are shown a film clip of robot abuse (see for instance Rosenthal-Von Der Pütten et al., 2014), or indicating behavioural intentions after interaction with a robot (Kahn Jr et al., 2012). The discussion about whether depictions of human-robot interaction, such as videos and texts, could be used as valid stimuli in HRI studies is ongoing. Robert Sparrow argued that such representations have sufficient moral value to serve as a test for the humans' virtue (Sparrow, 2017). Previous studies have shown that virtual representations of robots elicit more social behaviour (e.g. mimicking expressions, feelings of empathy, polite behaviour,

and physiological responses) than audiotapes or text (Rosenthal-von der Pütten et al., 2013; Slater et al., 2006), indicating that virtual robots, too, are recognised as social agents. J. Li (2015) conducted a meta-analysis on papers that studied the influence of agent embodiment on users' perception of the agent, and concluded that embodied robots elicited stronger behavioural and attitudinal responses than virtual agents. Several studies which had found no difference in behavioural and attitudinal responses for virtual agents and physical robots were missing in this analysis, however (for example Powers, Kiesler, Fussell, & Torrey, 2007; Reichenbach, Bartneck, & Carpenter, 2006). More recent studies found that the perception of and response to virtual agents was identical to embodied robots (Thellman, Silvervarg, Gulz, & Ziemke, 2016; Wullenkord, Fraune, Eyssel, & Šabanović, 2016). More specifically, Thellman et al. (2016) found that social presence (i.e., whether the robot was perceived as a social actor that manifests humanness (Lee, 2004)) rather than physical presence predicts social influence of a robot. In their experiment, social presence was not influenced by the physical embodiment of the robot. At the same time, Keijsers, Kazmi, et al. (2019, see also Chapter 6) found that robot embodiment had an effect on people's willingness to administer punishments: embodied robots got less severe punishment than their virtual replica. The discussion is, in other words, still ongoing. While studies seem to confirm that virtual agents do elicit social responses, the question remains if these are as intense as they would have been with an embodied robot. That being said, there is little doubt that virtual representations of robots can elicit an emotional response.

This was demonstrated as well by the public response to a video of a man kicking a robot dog. In February 2015, Boston Dynamics published a video of its quadruped robot "Spot". Employees kicked the robot in order to demonstrate the robot's capacity to regain its balance[1]. The video went viral and sparked discussions about the morality of the demonstrated behaviour (Sparrow, 2016), with many commenters perceiving the kicks to be abusive (Darling, 2015). In other videos, Boston Dynamics employees used a hockey stick to remove a box from the grip of Atlas, a humanoid robot[2]. The intention was to demonstrate Atlas' capacity to dynamically track and grip a box. Many viewers of the video however considered it to be mean-spirited teasing behaviour. Boston Dynamics has since included a disclaimer to their robot videos to assure viewers that the behaviour "does not irritate or harm the robot"(Boston Dynamics, 2018).

It seemed therefore feasible to study how people responded to robot abuse by collecting their responses to video recordings of more extreme cases of robot abuse than would be possible to set up in a lab experiment. Comparing or even benchmarking these responses to how people react to humans being exposed to the same abusive behaviour remained, however, more problematic. Up until now, no stimuli were available that would convincingly show the exact same abusive behaviour towards a robot and towards a human. Comparison studies were therefore often constrained to text stimuli.

[1] https://youtu.be/M8YjvHYbZ9w
[2] https://youtu.be/rVlhMGQgDkY

2.1.2 Current study

In June 2019 a mock Boston Dynamics video was released, in which an Atlas robot was shown as it performed a number of tasks while a human engineer deliberately attempted to sabotage them. These sabotaging behaviours got gradually more aggressive, until the robot turned and attacked the "bullying" human[3]. The robot in this video was CGI rendered; its motions had been captured through a human in a tracksuit. As a result, there were two versions of a video with identical abusive behaviours: one video where the victim was a human in a track suit, and a second version where the victim was a robot. This unique footage allowed us to compare the perceived morality of the exact same abusive behaviour when carried out towards a human versus a robot. See Figure 2.1 for a side-by-side comparison of the same frame for the human and the robotic agent.

Experiment I showed participants 14 instances of abusive behaviour towards either the robot or the human agent, and measured how morally (un)acceptable these behaviours were perceived to be. After the 14 videos that showed aggression towards the agent, two additional video clips were shown where the agent started fighting back (i.e. reactive aggression). Thus, the moral acceptability of reactive aggression to the group that just abused the agent was assessed. After the 16 video clips, participants assessed the agent's capability to think and feel (i.e. mind attribution).

Research questions

The research questions are as follows:

1. Is abusing a robotic agent seen as more morally acceptable than abusing a human?
2. Is reactive aggression more acceptable when it comes from the human agent than from the robotic agent?
3. If abusing one agent is seen as more acceptable than the other, is this difference in acceptability due to a different perception of how abusive the behaviour was?
4. Is mind attribution to the agent related to the moral acceptability of abusing it?

Hypotheses

These research questions led to the following hypotheses to be tested.

1. Based on the Media Equation theory (Nass et al., 1994; Reeves & Nass, 1996) as well as empirical evidence that viewing similar (although not identical) abusive behaviours towards robots and humans elicits similar neurological responses (Rosenthal-Von Der Pütten et al., 2014), it was expected that the abuse of the robotic agent would be considered equally acceptable as the abuse of the human agent.

[3] `https://youtu.be/dKjCWfuvYxQ`

2. Since it was expected that abusive behaviour to both agents would be seen as equally unacceptable, it was furthermore expected that there would be no difference in how acceptable reactive aggression from the agent was seen.

3. (a) Considering the findings from research on mind perception and empathy (Urquiza-Haas & Kotrschal, 2015), it was expected that for both the human and the robotic agent mind attribution would negatively correlate with how acceptable the abusive behaviour was rated.

   (b) If this relationship would be moderated by agent type, it was expected that the strength of the correlation would be affected; but not that the correlation would vanish for either agent nor that its direction would be inverted.

2.2 Method

2.2.1 Participants and Design

The experiment followed a single factor (agent: human or robot) between-subject design. Participants watched 14 videos in randomised order, in which an agent was exposed to various types of abuse. After each video, they rated the behaviour shown in the video clip for moral acceptability. These 14 videos were then followed by two videos in which the agent fought back. These two videos as well were each rated for moral acceptability. Finally, participants filled out questionnaires on mind attribution to the agent, individual tendency to anthropomorphise, and affinity with technology.

The dependent variables were: perceived acceptability of the videos where the agent was abused, perceived acceptability of the videos where the agent fought back, and mind attribution to the agent. Individual tendencies to anthropomorphise and affinity with technology were used as randomisation checks. This study was approved by the University of Canterbury Ethics board under the reference HEC 2019/30/LR-PS.

166 participants were recruited from Amazon Mechanical Turk (MTurk). Previous studies have indicated that data collected via MTurk is of equal quality to on-campus recruitment or participant data from forums (Bartneck, Duenser, Moltchanova, & Zawieska, 2015; Simons & Chabris, 2012), with internal motivation rather than monetary reward being the main motive for participating (Buhrmester, Kwang, & Gosling, 2011). Participants received 1 USD for their participation, which is in line with MTurk reimbursement custom. The survey took approximately 10 minutes to complete.

All participants were native English speakers and lived in the USA, UK, Canada, Ireland, Australia or New Zealand. All participants were Amazon Mechanical Turk Master Workers. These workers are being monitored by Amazon for their performance over time. Amazon explains that "Workers who have demonstrated excellence across a wide range of tasks are awarded the Masters Qualification. Masters must continue to pass our statistical monitoring to retain their qualification".

Of the 166 participants, nine reported being familiar with the video material; 30 had not seen the clip but thought the material was unrealistic. All these participants were

removed from the dataset, resulting in a final dataset of 127 participants. 51.18% (N = 65) were male; the mean age was 42.57 years (SD = 11.20; range = 25-72). 59 participants saw the human agent videos, while the other 68 were in the robotic agent condition.

2.2.2 Measurements

Moral acceptability of abuse

We measured the moral acceptability of the aggressive behaviour shown in the video clip through a single item, assessed after each video. The item stated *How (un)acceptable would you say the behaviour shown in the video is?* Participants could indicate their answer on a seven-point scale which consisted of the following answer options (see also Figure 2.2):

Forbidden	Unacceptable	Frowned upon	Discretionary	Suggested	Called for	Required

The terminology for the response options was taken from work on the dimensions of normative demand by Malle, Bello, and Scheutz (2019). Malle et al. validated a scale that held 13 points and ranged from prescriptions to prohibitions. To keep the scale readable for participants and avoid any formatting issues that could occur when an extensive scale would be displayed on a smaller screen (e.g. of a tablet or laptop) we reduced the number of items to 7 by omitting every other point on the original scale. Previous analyses have indicated that from 5 to 7 points on (depending on the covariance between the items) adding extra points to a scale does not alter the reliability of a scale (Lissitz & Green, 1975).

Since acceptability was measured with a single item only, its construct validity was assessed by correlating it to perceived violence, abusiveness, and intention to harm. Based on Cohen (1992), Pearson correlation coefficients of .6 or higher (i.e. a large effect size) were expected.

Violence, abusiveness, and intention to harm

Participants were asked to rate each video on three additional scales: how violent they thought the behaviour was, to what extent the behaviour had been intended to harm, and how abusive the behaviour was. Each item was answered on a 7-point scale ranging from "Not at all" to "Very much". See Figure 2.2 for a screenshot of one of the videos plus the four questions.

Individual differences in anthropomorphism

After having seen and rated all 16 videos, participants completed the individual differences in anthropomorphism questionnaire (Waytz, Cacioppo, & Epley, 2010), which measured their personal tendency to attribute different aspects of sentience and various emotions to a wide range of non-human entities (e.g. natural phenomena and animals). The original scale includes items that refer to mechanical entities as well (e.g. robots, cars, computers). For the current experiment, these were omitted since the experimental manipulation could

bias responses on those items. Individual differences in anthropomorphism was used as a randomisation check between the conditions. See Appendix A for the full questionnaire.

Mind attribution

Mind attribution to the agent was measured by means of the Mind Perception scale by Gray et al. (2007). This scale requires a person to indicate to what extent they believe an agent to be capable of eighteen different attributes (e.g., "Feeling afraid or fearful", "Understanding how others are feeling", and "Having personality traits that make it unique from others"). The score for each item is on a five-point range, from "not capable" to "extremely capable". See Appendix A for the full questionnaire.

Control questions

Finally, two control questions at the very end of the survey were included: (1) Have you seen this particular video before?; and (2) How authentic were the movie clips (on a seven-point scale, ranging from *Obviously not realistic (animated)* to *Clearly realistic*)? Participants who responded with "Definitely yes" or "Probably yes" to the first question, or the lower end (i.e. 1, 2, or 3) of the realism scale, were excluded from the analyses.

2.2.3 Video material

On 15 June 2019, Corridor Digital published a video in which the Boston Dynamics' Atlas robot was shown executing various attempts at picking up and carrying around a cardboard box, under supervision of a human engineer. In the video, another human engineer performs a variety of abusive behaviours towards the robot. These start with the type of behaviour Boston Dynamics shows its original videos, such as kicking the agent or using a hockey stick to interfere with the agent grabbing the box. Over time however, the behaviours get increasingly abusive, and peak with the human engineer shooting the robot. Eventually the robot starts fighting back and the roles of bully and victim get reversed. The robot forces the engineers to carry the boxes by holding them at gun point.

This special effects video was extremely well done and fooled nearly everyone to believe that an actual Atlas robot was used instead of a computer-generated model. Corridor Digital used motion tracking of a human actor to capture the behaviour and mapped a digital Atlas robot onto the movements to create the animation. Upon request, they kindly shared both the motion-capturing footage showing the human actor, and the special effects video with the Atlas robot (see Figure 2.1 for a side-by-side comparison between the same frame from the unedited video and the complete special effects video).

Each of the two versions was cut into 16 video clips. 14 of those depicted abusive behaviour towards the agent, two showed the agent responding with aggression to the human engineers. The 14 abusive videos showed a wide range of aggressive behaviour. Three scenes did not include any physical abuse but instead contained verbal abuse, such as "You are completely useless!" The other 11 video clips showed physical abuse or taunting.

The resolution of the videos was reduced from 3840×2160 pixels to 1280×720 pixels to ensure fast playback on mobile devices. Video playback speed was tested before the

experiment was run, and could no streaming delays were detected. The videos are available as supplementary material to this thesis.

2.2.4 Procedure

Prospective participants could select the task in MTurk to read a short description of the study. If they decided to participate, they were directed to a Qualtrics survey page. After informed consent was provided and demographics (age, gender) were assessed, the participants were randomly assigned to one of the two agent conditions (human or robot). They were then given instructions to ensure that their audio playback was working. Within each condition, they watched the videos in a randomised order. After watching each video, the participants provided responses on the dependent measures, before moving to the next video (see Figure 2.2). After the main experiment, the participants filled out the individual differences in anthropomorphism scale and the two control questions. Then they were thanked for their time, offered a debriefing and given the reimbursement code which they could use to claim their reward at MTurk. The entire experiment took about 10 minutes to complete.

2.3 Results

2.3.1 Preliminary analyses

Exclusion of participants

39 participants were excluded due to reporting that the video material was unrealistic (N = 34) and/or reporting that they had already seen the material before (N = 9). Of the participants who deemed the material unrealistic, 23 saw the human agent video and 11 saw the robotic agent video. Of the nine participants who reported they'd seen the video before, six were in the robotic agent condition. Participants who were excluded had lower levels of individual tendency to anthropomorphise, M(SD) = 2.31(*.77*), than participants who were left in the analysis, M(SD) = 2.63(*.81*), t(75.94) = -2.28, p = .026. Excluding those participants thus may have introduced a bias to the results. We discuss the potential biases in more detail in the limitations section. We chose to report the results for the dataset with those participants excluded, but ran the same series of tests for the original (full) dataset. If the findings diverged, that is, if a significant effect became insignificant or vice versa, we reported both the results on the full dataset as well.

Confound check for an interaction effect between agent and video material

Before collapsing the fourteen measurements of acceptability of abuse, a confound check was performed. Considering how the abusive behaviours covered a wide range of bullying behaviours, the possibility exists that one or more specific abuses would be considered unacceptable for one agent, but not the other. Variability between the fourteen abuses is to be expected, but agent-specific variability would mean that hypothesis 1 (no differences between how acceptable abuse of a human versus robotic agent is seen) would have to

Please watch the video carefully and then answer the following questions:

How (un)acceptable would you say the behaviour shown in the video is?

Forbidden	Unacceptable	Frowned upon	Discretionary	Suggested	Called for	Required
○	○	○	○	○	○	○

	1 - Not at all	2	3	4 - Neutral	5	6	7 - Very much
How violent would you rate the behaviour in this video?	○	○	○	○	○	○	○
To what extent do you think the behaviour was intended to harm?	○	○	○	○	○	○	○
How abusive would you rate the behaviour in this video?	○	○	○	○	○	○	○

Figure 2.2: Screenshot of one of the videos plus the questionnaire.

be rejected. When the measures are collapsed into a single index of abusive behaviour however, this difference might disappear, thus creating a confound.

Thus, we tested for an interaction between each video and the agent on each of the four measurements. That is: moral acceptability, perceived violence, abusiveness, and the intention to hurt. Note that the objective of this test was explicitly not to look for main effects. Any differences in perceived violence between video A and video B, or in ratings of abusiveness between robotic and human agents, were not considered. Instead, the main point of interest were the interaction effects. For example, whether people thought that the behaviour in video C was way more hurtful than the behaviour in all other videos, but only if the agent was a robot. Without such an interaction the 14 measurements could be aggregated into a single measure of how morally acceptable overall mistreatment of the agent was perceived. The 14 videos thereby became a representative index of abusive behaviour.

For each of the four measurements, a 14 $\times$ 2 mixed effects ANOVA with the measurement as dependent variable, the video ID as within subject factor and the agent as between subject factor was carried out. None of the interaction terms was significant, $\chi(13) < 14.80$, $ps > .320$, indicating that it was possible to average the 14 into a single "moral acceptability" measure.

Reliability tests

Internal consistency was high for the individual tendency to anthropomorphise scale, $\alpha = .83$. The mind attribution scale too had high internal consistency, $\alpha = .98$. The scales were therefore deemed reliable (Cronbach, 1951).

Randomisation tests

Individual tendency to anthropomorphise did not differ between conditions, M(SD) = 2.61(*.69*) and 2.64(*.91*) for human and robotic agents respectively, $t(125) = -0.160$, $p = .873$. Gender was evenly distributed between the conditions, $\chi^2(1) = 2.79$, $p = .09$. For the full dataset however, gender was not evenly distributed, with 38 out of the 82 participants in the human agent condition being male (45.24%) and 51 out of 82 being male (62.20%) in the robotic agent condition. As gender was not related to the dependent variable nor the mediator variable, $\chi^2(1) < .89$, $p > .344$, this imbalance was not considered problematic.

Pearson correlation between acceptability and violence, abusiveness, and intention to harm

Pearson correlation coefficients were calculated between acceptability of robot abuse on one hand, and perceived violence, perceived abusiveness, and intention to harm on the other. The correlation coefficients all exceeded the benchmark for a large effect (Cohen, 1992), $\rho > .637$, $p < .001$. As a result, assessing acceptability of abuse with a single-item measure is not considered problematic for the construct validity.

2.3.2 Main analyses

To answer our first two research questions, two independent sample t-tests are conducted. In each test, "acceptability of aggression portrayed in the video" is the dependent variable, and agent (human or robotic) the independent variable. The first t-test concerns the 14 videos with aggression towards the agent; the second t-test covered the reactive aggression videos. In addition, a TOST procedure was carried out to test for equivalence between the two conditions (Richter & Richter, 2002). The smallest effect size that was expected to be detected with 80% power was used to set the boundaries. This effect size was calculated by means of the G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007).

A 2×2 mixed ANOVA with factors "agent" (human versus robot; between participants) and "aggression" (unprovoked i.e. by the bullies versus reactive i.e. by the agent; within participants) was considered but would likely have resulted in biased outcomes for the "aggression" main effects, as the unprovoked videos were both more diverse and intense (ranging from verbal abuse to lethal aggression) and higher in number.

Acceptability of aggression towards agent

Participants in the human agent condition rated the videos as equally acceptable, M(*SD*) = 2.40(*.49*), as participants in the robot agent condition, M(*SD*) = 2.53(*.80*); $t(125) = 1.10$, $p = .275$. The TOST equivalence test confirmed that acceptability ratings between the two conditions were equivalent $t(113.43) = 1.74$, $p = .043$, given equivalence bounds of -0.334 and 0.334 (on a raw scale) and an alpha of 0.05.

(For the dataset including the participants who regarded the video material unrealistic or indicated that they had seen the materials before, a marginal effect was found. Abuse of a robotic agent was rated as slightly more acceptable, M(*SD*) = 2.61(*.80*), than abuse of a human agent, M(*SD*) = 2.41(*.56*), $t(164) = -1.79$, $p = .076$. The TOST equivalence did not return significant, $t(144.18) = 1.026$, $p = .153$, given equivalence bounds of -0.302 and 0.302 (on a raw scale) and an alpha of 0.05.)

Acceptability of reactive aggression from agent

Participants in the human agent condition rated the reactive aggression as more acceptable, M(*SD*) = 3.74(*1.45*) than participants in the robotic agent condition, M(*SD*) = 2.99(*1.33*), $t(125) = -3.02$, $p = .003$.

Difference between agents: mediation

In order to answer our third research question, mediation analyses on the significant findings as indicated by the t-tests (if any) are conducted. In a mediation analysis, one tries to gain further understanding of a relationship between an independent variable (IV; here: agent) and a dependent variable (DV; here: acceptability of aggression) by including a third variable in the analysis (mediator; here: perceived abuse), which is hypothesised to be related to both the dependent and the independent variable. A mediation model proposes that (part of) the relationship between the IV and the DV is because the IV has

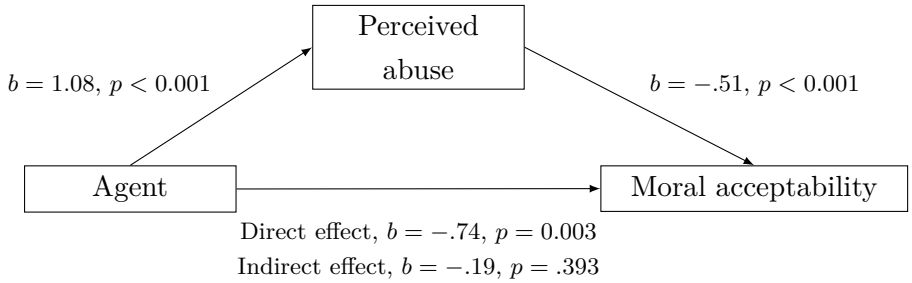


Figure 2.3: Mediation model showing the full mediation by perceived abuse on the relationship between agent and moral acceptability of reactive aggression (i.e. the agent fighting back). Note that the default for Agent was set at *human*, so that reactive aggression was seen as less acceptable and more abusive for the robotic agent.

an effect on the mediator, which in turn influences the DV. In the current experiment, this would mean that any relationship between agent and the acceptability of aggressive behaviour shown in the video would (partially or only) exist because the agent would have an effect on how abusive the behaviour was perceived to be; and how abusive the behaviour was seen to be influenced the moral acceptability of it.

Mediation analysis can only be performed on a significant relationship between an independent and a dependent variable. Since a significant effect of agent on acceptability had only been established for reactive aggression, acceptability of aggression towards the agent will not be considered for mediation analysis.

For the first step of the mediation analysis, acceptability was regressed on agent; as found in 2.3.2, this relationship was significant, b = -.74, $t(125)$ = -3.02, p = .003.

For the second step, a significant relationship between the independent factor and the mediator has to be established. Abuse was regressed on agent; this relationship, too, was significant, $b = 1.08$, $t(125) = 4.13$, $p < .001$. For the final step of the mediation analysis, acceptability was regressed on both agent and abuse. A full mediation occurred, with agent dropping as a predictor (b = -.19, $t(124)$ = -.86, p = .393) and perceived abusiveness taking over as the only significant predictor (b = -.51, $t(124)$ = -7.22, p <.001). Sobel's test confirmed the significance of this effect; Z = -3.59, $p < .001$. See 2.3 for the mediation model.

Mind attribution and acceptability

To test whether moral acceptability was dependent on mind attribution to the agent, a regression analysis was performed with acceptability of agent mistreatment being regressed on mind attribution to the agent. Mind attribution was significantly related to acceptability of agent mistreatment, b = -.09, $t(125)$ = -2.15, p = .034. The more a participant attributed the agent a mind, the less acceptable abuse was perceived to be.

Since mind attribution had been significantly different between the two types of agents, a one-way ANCOVA with agent as independent variable, mind attribution as covariate, and acceptability of mistreatment was ran to determine whether this relationship was independent of the type of agent. Agent type as well as an interaction term between agent

type and mind attribution was not significant, $ps > .36$. This confirmed that the relationship between mind attribution and acceptability of mistreatment was not influenced by agent type.

2.4 Discussion

Experiment I compared differences in acceptability between abuse of a human and abuse of a robotic victim. Markedly, the video materials that were used were both of exceptional quality (making it hard to recognise the robot for the CGI rendering it was), and showed the exact same bullying behaviour to either of the two agents. In addition, the materials covered a wide range of bullying behaviours. As a result, the materials used were both highly realistic and perfectly synchronised except for the agent depicted.

Three hypotheses were tested. Hypothesis 1, which stated that there would be no differences in the acceptability of abusive behaviour towards robots or humans, was confirmed. The participants considered mistreating a robot to be as immoral as abusing a human. While this may not automatically mean that the participants consider robots to be equivalent to humans in all respects and in all situations, it does show at least that bullying behaviour is considered immoral, no matter who the victim is.

Secondly, and in line with hypothesis 3a, perceived capability of the agent to think and feel was negatively related to how acceptable people found the abuse of the agent. This is also in line with previous research, where mind attribution has been related to empathy (see for instance Gray et al., 2007; Urquiza-Haas & Kotrschal, 2015). Moreover, hypothesis 3b suggested that if a moderation of this relationship by agent type would have been found, only the strength of the relationship (rather than direction) would have been affected. However, it was found that agent type did not moderate this correlation between mind attribution and moral acceptability of abuse. This implies that the abuse of the robot was seen as unacceptable for the same reason as the abuse of the human. This finding also undermines the alternative theory that participants considered bullying the robot morally unacceptable because they saw it as damaging property, while the human bullying would be seen unacceptable because it was perceived as actual bullying. If this had been the case, a moderation effect would have been found, where mind attribution only predicted moral acceptability for human agents whereas there was no relationship between mind attribution and acceptability of the abuse of the robot agent.

Finally, hypothesis 2 was rejected, as reactive aggression by the agent was not considered equally acceptable for each type of agent. Specifically, participants found it significantly more acceptable if the human started fighting back, than if the robot fought back. Mediation analysis showed that this was a consequence of the reactive aggression being perceived as more abusive when it came from a robot. This was in spite of the acts of responsive aggression being identical for the robotic and the human victim. This is not the first time that identical behaviour is perceived as more or less intentional depending on whether the actor is human or robotic. Thellman et al. (2016) compared perceived intentionality and control of an agent (either human or robotic) for a range of positive and negative behaviours. They found that positive behaviour is seen as more controllable and

intentional when carried out by a human than by a robot. Negative behaviour however was seen as more intentional when performed by a robot than by a human. What could have caused this asymmetry?

We speculate that robots could have been perceived as deserving protection from harm to the same extent as humans, but were not perceived to have the same right of self-defence. To our knowledge, there are only two HRI papers that relate to these findings: Kahn Jr et al. (2012) had children of various ages interact with a Robovie humanoid robot before it was locked away in a closet. Robovie protested against this treatment. The children were interviewed about a range of topics, including the robot's moral standing. For the oldest age group (the 15-year olds), an interesting pattern emerged: slightly more than half of those children thought it was wrong to hurt the robot by locking it away or eventually crushing it when it would be no longer needed. The vast majority, however, did not think the robot should be paid for a hard day's work or be granted the right to vote; and less than one in 10 thought the concept of owning and selling the robot was wrong. This pattern — a right to be protected from harm, but no right to autonomy — is surprisingly similar to what was found in our study. It mirrors the ethical view many have towards animal rights. While animals can be considered property and can even be killed, mistreatment is not allowed. Animal rights has therefore been proposed as a template for robot rights (Calverley, 2006).

A different perspective could be offered by a study on the trolley dilemma (Malle et al., 2015). In this moral dilemma, a trolley is rushing down the track at great speed, and will hit and kill four people if not side-tracked to a route where it will kill only one person. People have to choose between not taking any action, thus indirectly being responsible for the death of four people, or taking action and being directly responsible for the death of one person. In spite of the net saving of three lives when acting, most people find acting harder than not taking any action. However, Malle et al. (2015) found that robots were more strongly expected to make a rational choice and more strongly blamed if they went with the emotional solution. On the contrary, humans were blamed more if they chose to divert the train. In this light, the robot's reactive aggression could be considered more wrong as we expect robots to be more rational and less affected by emotion when choosing to act, while emotion is expected to play a role in human moral decision making. In the current experiment, this means that participants would have expected the robot to "keep its cool" rather than counterattack with the strength that it did.

A third possible explanation could be that participants interpreted the robot's reactive aggression as more intimidating than the human's reactive aggression. A robot fighting back might be considered as dangerous, as robots are often portrayed in public media as a potential threat to mankind. The trope is that robots raise up against their masters and enslave humanity (Bartneck, 2013; Złotowski, Yogeeswaran, & Bartneck, 2017). This trope of robots shaking off the yoke of servility and subduing mankind may have been triggered in participants viewing the robot fight back against the bullies. While the aggression was exactly the same as the aggression expressed by the human agent, the robot's aggression would be interpreted as more dangerous as it might generalise to all humans, whereas the

human's aggression would be restricted to fighting off the bullies.

Previous research had established that people can feel empathy for robots when the robot is mistreated (e.g. Darling, Nandy, & Breazeal, 2015; Riek, Rabinowitch, Chakrabarti, & Robinson, 2009; Rosenthal-Von Der Pütten et al., 2014). Experiment I extrapolated on these findings in two ways. Firstly, it was shown that people consider robot bullying as unacceptable as human bullying. Secondly, it was shown that this unacceptability is linked to mind perception in the victim, regardless of whether this victim is human or robotic. To our knowledge, this is the first experiment to compare human bullying to robot bullying in terms of morality.

This is important because although empathy and morality are related, they are not necessarily the same. People could feel sorry for the robot yet see robot bullying as justified. Moreover, Experiment I provides initial evidence that mind attribution might play a role in empathy with robots and robot bullying.

2.4.1 Limitations

There are a few limitations to Experiment I that should be noted. Firstly, the videos showing the human actor were used to motion capture the movement of the digital Atlas robot. For this purpose, the human actor moved in a rigid way that's stereotypical of robots, which of course deviates from natural human movement. At one hand, this may have been a bit confusing to the participants. On the other hand, it further reduced any differences between the human and robot agent condition. For example, a different movement pattern for the human agent may have introduced bias through providing (implicit) information on the human being hurt. We believe that the movements of the actor were sufficiently plausible movements for a human to not cause any confusion in participants as of whether the agent was in fact human. This also showed in the mind attribution scores, which were high (between "capable" and "very capable") for the human agent.

A second limitation of this experiment is that the Corridor Digital company added some small special effects to the robot video, in particular in the section in which one of the engineers fires a hand gun at the robot. In the robot video there is additional nozzle fire, smoke and impact indicators. All these effects are subtle and the focus of each scene is the interaction between agent and bullies rather than these minor special effects. We believe that these slight differences do not significantly impact the similarities of the videos. Most videos were identical between the human and the robot condition.

We excluded participants who considered the videos as unrealistic from our statistical analyses. This might have introduced a small bias since these participants also had a slightly different tendency to anthropomorphise. Still, including them would have also introduced another bias, namely that of participants who did not suspend their disbelief. We believe that the later would have been the stronger bias and hence our decision to exclude the participants was the better choice. To have a better insight into how the exclusion may have influenced the results, we ran the same analyses once more with the complete dataset. No new significant results emerged and no previously significant results turned insignificant, with the exception of the TOST.

2.4.2 Conclusion

While previous work considered empathetic concern towards robots in relation to robot abuse (e.g. Darling et al., 2015; Riek et al., 2009; Rosenthal-Von Der Pütten et al., 2014), the current experiment looked at moral acceptability of abuse and also compared the acceptability of robot abuse to acceptability of human abuse. Interestingly, the results suggested that people do not consider robot bullying more acceptable than human bullying. The link between mind attributed to the victim and perceived acceptability of abuse was similar for robotic and human victims alike.

These results confirm that people view robot abuse as bullying. In addition, they provide insight to factors that may prove to be central to robot bullying behaviour through establishing a link to mind attribution. They strongly suggest that parallels can be drawn between robot abuse and human bullying, which is valuable information for researchers in HRI.

Chapter 3

Experiment II: Robot sentience and acceptability of mistreatment

3.1 Introduction

Experiment I (described in Chapter 2) established that people do not perceive robot bullying as significantly different from human bullying. Moreover, it was found that mind attribution was related to perceived acceptability of bullying behaviour. However, this relationship was correlational, since mind attribution had not been directly manipulated. As a result, no causal inferences could be made about the direction of this relationship — individuals who have more stringent moral standards might also have an enhanced tendency to attribute mind to others, for example. Thus, in order to test for a causal relationship between mind attribution and moral acceptability, under Experiment II two experiments were conducted that manipulated the perception of a robot's capability to think and feel and subsequently measured the effects on acceptability of robot bullying, as well as people's own tendency to bully the robot.

3.1.1 Literature

In Experiment II, we raise the question of how robot mind attribution influences perceived right to moral treatment and willingness to bully the robot. Previous research has already found a relationship between robot anthropomorphism and willingness to harm. Higher robot anthropomorphism, manipulated by making the robot look more human-like, has been linked to a greater willingness to protect it from harm (Riek et al., 2009). Anthropomorphising a simple bug robot through providing it with a background story increased participants' hesitation before harming it (Darling et al., 2015). Similarly, functional robots allegedly were less often the target of abuse when they malfunctioned if they had been given a name, which would have increased the robot's anthropomorphic qualities (Darling, 2015). These examples, however, employed basic ways of enhancing anthropomorphism and did not directly target perceived mind as an indicator of anthropomorphism. Briggs and Scheutz (2014) did manipulate mind attribution, and found that when a small humanoid robot expressed its distress, participants' behaviour was indeed

affected. People were less likely to insist that a NAO robot should follow the command to topple over a tower it had just painstakingly built when the robot protested in an emotional way than when it stayed silent (Briggs & Scheutz, 2014). In Bartneck, Van Der Hoek, et al. (2007), participants hesitated longer before switching off an expressive iCat robot when their interaction with it revealed it as intelligent and agreeable, than when it communicated in a manner that was either intelligent, or agreeable, or neither.

In social robotics, robots are often programmed to display social cues, for example through expressing emotions and non-verbal behaviour. Since these cues would increase the extent to which the robot is perceived as possessing a mind of its own (Złotowski et al., 2014, 2018), one would expect that including such cues in a robot's behaviour increases empathy with the robot and decreases participant's willingness to harm it. Yet this relationship appears to be complicated.

A number of studies suggest that a higher degree of social behaviour by a robot may not always be related to a lower willingness in participants to harm it (Horstmann et al., 2018; Nomura et al., 2016; Tan et al., 2018). In contrast to the iCat study by Bartneck, Van Der Hoek, et al. (2007), Horstmann et al. (2018) found that a functional robot that protested against being switched off resulted in longer hesitation and lesser inclination to switch it off than a protesting robot with social behaviour. Tan et al. (2018) measured participants' willingness to intervene as a confederate verbally and physically bullied a small Cozmo robot. There was a marginal trend where participants were more likely to discourage mistreatment of the robot if it did not display any emotional behaviour throughout the experiment, versus when it celebrated successes and mourned losses. Nomura et al. (2016) interviewed children who bullied the anthropomorphic Robovie robot in a shopping mall. Most children saw the robot as human-like rather than machine-like and about half saw the robot as capable of perceiving its environment. Yet neither observation had stopped them from physically and verbally bullying the robot.

How to explain these inconsistent findings? The argued link between social cues and willingness to harm consists of three propositions. First of all, it assumes that social behaviour by a robot would enhance mind attribution. Emotional cues have indeed shown to increase the attribution of mental capabilities (Złotowski et al., 2014), comprehension skills (Briggs & Scheutz, 2014), and ability to experience emotion (Tan et al., 2018). This proposition thus seems to be supported by empirical evidence.

The second and third propositions of the argument are that mind attribution enhances empathy, and that empathy in turn decreases willingness to harm. There has been some debate on whether empathy and mind perception are different concepts in the first place (Premack & Woodruff, 1978; Whiten & Byrne, 1991, e.g.) but fMRI studies have found that attributing a mind to others activates different neural networks than empathy (Hein & Singer, 2008).

The second and third proposition furthermore are at the core of dehumanisation theory (Haslam, 2006; Haslam, Loughnan, et al., 2008). Dehumanisation theory asserts that lower mind attribution to another person allows someone to disregard the negative consequences of their behaviour for that person, as the reduction in perceived mental capabilities renders

the victim less deserving of empathy. This was to some extent collaborated by research by Gray et al. (2007), who found that mind perception was related to moral status. Experiment I (Chapter 5) as well found a correlation between mind attribution and perceived acceptability of abusive behaviour.

However, several studies suggest that empathy may not automatically lead to more protective behaviour (Bartneck, Van Der Hoek, et al., 2007; Horstmann et al., 2018; Tan et al., 2018). Horstmann et al. (2018) reports that participants experienced higher stress levels while switching off a protesting robot when they liked the robot better, but did not take any longer to turn it off. Bartneck, Van Der Hoek, et al. (2007) did find that participants hesitated longer to switch off an agreeable robot, but also noted that all participants eventually turned off the robot whether it was smart and agreeable or not. In the study by Tan et al. (2018) on whether participants intervened when they saw a Cozmo robot being bullied, participants who felt more strongly that the robot was mistreated did not display greater willingness to intervene.

Rosenthal-Von Der Pütten et al. (2014) compared participants' brain activation patterns in response to viewing the abuse of a cardboard box, a robot, and a human. They found that in the questionnaires people attributed equal levels of emotion to the human and the robot, and reported feeling the same amount of empathy towards the robot and the human when it was mistreated. In contrast, fMRI scans showed greater activation in participants' right putamen when watching the human being mistreated than when watching the robot being mistreated. This area has been associated with empathy and emotional distress (Rosenthal-Von Der Pütten et al., 2014). Interestingly, areas related to emotional processing and perspective taking were equally active when participants watched the robot or the human. Thus, self-reports and fMRI results suggest similar levels of mind perception to the robot and the human. However, when it came to empathy, people reported there was no difference while their brain activation suggested otherwise.

A potential explanation for the apparent contradiction between self-reports and areas of brain activation was provided by Urquiza-Haas and Kotrschal (2015). They proposed that empathy is the result of a cognitive dynamic, where implicit processes cause the emergence of early evaluations, and deliberate, reflective processes shape and nuance this evaluation. While early evaluations of robot bullying may elicit an initial empathetic response (which would be less intense than a response to human bullying), this response could then be enhanced or suppressed by cognitive appraisal of the behaviour the robot is exposed to.

It has been shown in human-human interaction that conscious processes can influence automatic responses (Liepelt & Brass, 2010; Liepelt, Cramon, & Brass, 2008). An adaptation of this research to the field of HRI investigated the influence of robot autonomy on automatic social processes by means of assessing a Social Simon task (Stenzel et al., 2012). In human-human interaction, the mere presence of another human decreases performance on this otherwise simple task; similar to the Stroop effect (Stroop, 1935), this is not consciously controlled. The robot's appearance and behaviour in this experiment were identical through the conditions, but half the participants were told upfront that

the robot was biology-inspired and had autonomous behaviour. The other half was told the robot was machine-like and deterministic; thus manipulating intentionality (one of the components of mind attribution). The Social Simon effect was observed only for the intentional robot. This proves that an explicit instruction can affect automatic behaviour (Stenzel et al., 2012).

3.1.2 Current studies

Research questions

The current research explored the following research questions in two experiments:

1. Does telling people that a robot possesses a mind, i.e., is capable of experiencing emotions and cognition, affect how unacceptable they find robot bullying?

2. Do emotional cues that imply the robot has a mind affect how unacceptable people find robot bullying?

3. Does explicit information that a robot does not possess a mind change the influence of implied mind on how unacceptable people find robot bullying?

4. Does telling people that a robot possesses a mind reduce their willingness to publicly humiliate it?

Experiment II.A was vignette-based and addressed the first three research questions. Robot sentience was manipulated in two ways: through having the robot display emotional cues, and through telling participants that the robot could think and feel. Participants then indicated how unacceptable they considered varying bullying behaviours towards the robot.

In Experiment II.B, the first and the last research question were addressed. Participants interacted with embodied robot that was introduced as either capable or incapable of thinking and feeling. They then indicated how unacceptable they considered bullying this robot; and were offered an opportunity to humiliate the robot they had just interacted with.

Hypotheses

The following hypotheses were formulated:

1. For Experiment II.A, it was expected that

   (a) explicit briefing that the robot possessed a mind would make people less accepting of robot bullying.

   (b) emotional cues by the robot that suggested it had a mind would decrease how acceptable people rated robot bullying.

2. For Experiment II.B, it was expected that

(a) the findings of Experiment II.A would be replicated in a setting with an embodied robot. That is, it was hypothesised that informing participants that the robot they were going to interact with was capable of thinking and feeling, would reduce how unacceptable they rated abuse of that robot post-interaction.

(b) participants would be more likely to humiliate the robot after being told it did not possess a mind.

The experiments have been reviewed and approved by the Human Ethics Committee at the University of Canterbury, under the reference HEC2019/47.

3.2 Experiment II.A

Experiment II.A was an online scenario-based study that followed a 2 (robot introduction: not possessing mind versus possessing mind) × 3 (robot response to mistreatment: no response, non-emotional response, emotional response) between participant design. Unacceptability of robot bullying as described in the scenario was the dependent variable. Participants' general tendency to anthropomorphise and affinity with technology were assessed in order to check whether they were similarly distributed across the six conditions.

3.2.1 Method

Participants

Participants for Experiment II.A were recruited using Amazon Mechanical Turk (MTurk), an online platform for data collection. Previous studies have indicated that data collected via MTurk are of equal quality to data collected through on-site recruitment or participant data from forums (Bartneck et al., 2015; Simons & Chabris, 2012), with internal motivation rather than monetary reward being the main motive for participating (Buhrmester et al., 2011). We restricted participation to participants residing in English-speaking countries (i.e. USA, Canada, Australia, New Zealand, United Kingdom, or Ireland) and accredited with Master status, i.e. with a low incidence of work being rejected.

A total of 193 people participated in Experiment 1. After having recruited the first 66 participants, it became apparent that basic demographics, gender and age, had not been assessed. Therefore, these demographics were included for the remainder of the data collection. Of the 126 participants who completed the survey after the error had been detected, 53.97% were female, with a mean age of 41.65 years (SD = 11.42). In return for their participation, workers were reimbursed with .90US$, in accordance with MTurk reimbursement custom.

Procedure

Prospective participants could read a short description of the study in MTurk. If they decided to participate, they were directed to a Qualtrics survey page, where they provided informed consent and reported their age and gender.

Subsequently, participants were randomly assigned to one of the six conditions. Depending on the condition, participants were presented with a vignette in which the robot was introduced as either possessing a mind (i.e. capable of various emotional and cognitive experiences) or not possessing a mind. The vignette further described a human-robot interaction between a participant and the robot which included bullying of the robot. Depending on the experimental condition, the robot responded to the bullying in a non-emotional way, in an emotional way, or not at all.

Finally, participants completed the survey. First, they indicated how morally unacceptable the bullying as described in the vignette. Subsequently, they completed the individual tendency to anthropomorphise and the affinity with technology scales. Then participants were thanked for their time, debriefed, and reimbursed.

Materials

Vignettes Vignettes constituted an introduction of the robot, a description of the robot being bullied, and a description of the robot's response to the bullying. In order to create the six conditions, nine vignettes were constructed: two different introductions (of the robot possessing and not possessing a mind), four different robot bullying scenarios, and three different robot responses (no response, non-emotional response, emotional response). See Table 3.1 for the respective introductions.

The four mistreatment descriptions were created to cover a wide scope of robot bullying. They described an interaction between the robot and a participant, where the participant had behaved in an aggressive or impolite way towards the robot: either playing around with the robot's energy supply; verbally abusing it; rejecting the proposal from the robot to split a monetary reward evenly and keeping all the money for themselves; or switching off the robot in spite of the robot asking to be left on sleep mode since switching it off would result in the robot losing awareness.

There were $2 \times 4 \times 3 = 24$ possible vignettes to which participants were randomly assigned. However, the interaction descriptions were not included as an independent variable as they were expected to have no effect on the dependent variable. A manipulation check was carried out to confirm this; see 3.2.2.

Measurements

Unacceptability of bullying Unacceptability of bullying was measured through five items, each scored on a 7-point Likert scale. The items concerned how opposed the participant was to treating the robot like it was treated in the vignette; if they considered the treatment as described acceptable; if they would intervene if they were to witness such treatment of a robot; how important it was to protect a robot like the one in the vignette from being treated like it was; and in general, how important it was for a robot like the one in the vignette to be treated humanely. The items were scored in such a way that a higher score indicated lower rated acceptability of robot bullying.

Table 3.1: Robot introductions manipulation. Left: introduction of the robot not possessing a mind. Right: introduction for the robot possessing a mind

Introduction not possessing a mind	Introduction possessing a mind
This robot is programmed to appear to be a social being: it is capable of processing, interpreting and calculating an emotional response to its environment. It can store and retrieve names and faces, so that it will state the name of people it has seen before out loud. It can also respond to prespecified commands, and will update its behaviour scheme to mimic an upset or angry response when given certain prompts. It has distance and depth sensors that prevent it from colliding with objects or people and falling off the stairs. All these behaviours give it the appearance of being conscious. The robot has starred in a few of our demos before, although it can't remember this. Recently it made its appearance in its first experiment. However, being a robot, it did not feel excited or nervous.	This is a social robot with his very own personality: it is capable of processing, interpreting and emotionally responding to its environment. It can remember names and faces, and will recognise people it met before. It can understand different commands, and will change its mood depending on how it is treated - for example, it will get upset when mistreated and happy after being told it did well. Moreover, it is aware of its surroundings, so that it can avoid bumping into objects or people and throwing itself off the stairs. The robot is proud to have starred in a few of our demos before, and recently made its appearance in its first experiment, for which it was very excited and a little nervous.

Individual differences in anthropomorphism questionnaire Individual differences in anthropomorphism were measured with a questionnaire from Waytz, Cacioppo, and Epley (2010), although the questions that targeted anthropomorphic qualities of technology (i.e. computers, cars and robots) were taken out since those would likely be affected by the introduction manipulation. The resulting questionnaire consisted of 10 items. Participants were asked to indicate on a 10-point Likert scale to what extent they thought different animals and natural phenomena have mental and emotional responses (e.g. "To what extent does the environment experience emotions?", "To what extent does the average insect have a mind of its own?"). See Appendix A for the full questionnaire.

Individual differences in anthropomorphism were assessed in order to check whether participants had similar trait anthropomorphism across the different conditions.

Affinity with technology The affinity with technology scale measures to what extent people are comfortable around, and eager to learn about technology. Like the individual differences in anthropomorphism, affinity with technology was assessed in order to check whether participants in different conditions had similar affinity with technology, as it was expected that differences between the conditions on this measure could bias the results.

Affinity with technology was measured with a questionnaire taken from Neyer, Felber, and Gebhardt (2012) and translated from German to English. Participants' individual affinity with technology is measured through their agreement with eight statements (e.g. "I am very curious about new technological developments") on a 5-point Likert scale (ranging from "not at all descriptive of me" to "extremely descriptive of me"). See Appendix A for the full questionnaire.

Table 3.2: Mean scores (*SD*) for age, trait anthropomorphism, and affinity with technology per condition

	No mind condition		
	no response	non-emotional	emotional
Age	50.06(*10.13*)	41.00(*12.39*)	38.83(*9.29*)
Percentage male	55.56%	21.74%	50%
Trait anthropomorphism	4.12(*1.59*)	4.54(*1.78*)	3.77(*1.77*)
Affinity with technology	3.91(*.77*)	3.73(*.93*)	4.05(*.94*)
Unacceptability robot bullying	2.71(*.98*)	3.01(*.97*)	3.16(*1.01*)
	Mind condition		
Age	41.28(*11.15*)	39.57(*12.86*)	40.65(*10.11*)
Percentage male	45.83%	33.33%	76.47%
Trait anthropomorphism	3.75(*1.46*)	3.90(*1.57*)	3.85(*1.80*)
Affinity with technology	3.70(*.92*)	3.96(*.80*)	3.83(*.70*)
Unacceptability robot bullying	3.02(*.98*)	3.39(*1.10*)	3.42(*1.16*)

Explicit Mind Attribution Explicit mind attribution to the robot was measured in order to perform a manipulation check on the robot mind attribution manipulation. The mind attribution questionnaire was taken from Gray et al. (2007) and adapted so that the questions explicitly referred to the robot from the introduction. The questionnaire measures to what extent the robot is capable of experiencing 18 different emotional and cognitive states, using a 5-point Likert scale that ranges from "very incapable" to "very capable". See Appendix A for the questionnaire.

3.2.2 Results

Preliminary analyses

Reliability Cronbach's alpha was computed for the unacceptability of robot bullying scale, as well as the individual differences in anthropomorphism questionnaire and the affinity with technology questionnaire. Internal consistency was high; $\alpha = .88$ for unacceptability of robot bullying, $\alpha = .85$ for anthropomorphism, and $\alpha = .88$ for affinity with technology. The questionnaires and scale were thus deemed reliable (Cronbach, 1951).

Bullying scenario confound check A 4×1 ANOVA with the four bullying scenarios as independent variables and unacceptability of bullying as dependent variable confirmed that the scenario did not influence unacceptability scores, $F(3, 189) = 1.01$, $p = .389$. The bullying scenarios thus could be excluded as a factor, as intended.

Randomisation checks A 2×3 ANOVA with the robot introduction and robot response as independent variables and age as dependent variable indicated that age was not equal across the conditions. More specifically, there were main effects for the introduction manipulation ($F(1, 121) = 6.56$, $p = .012$) as well as the robot response ($F(2, 121) =$

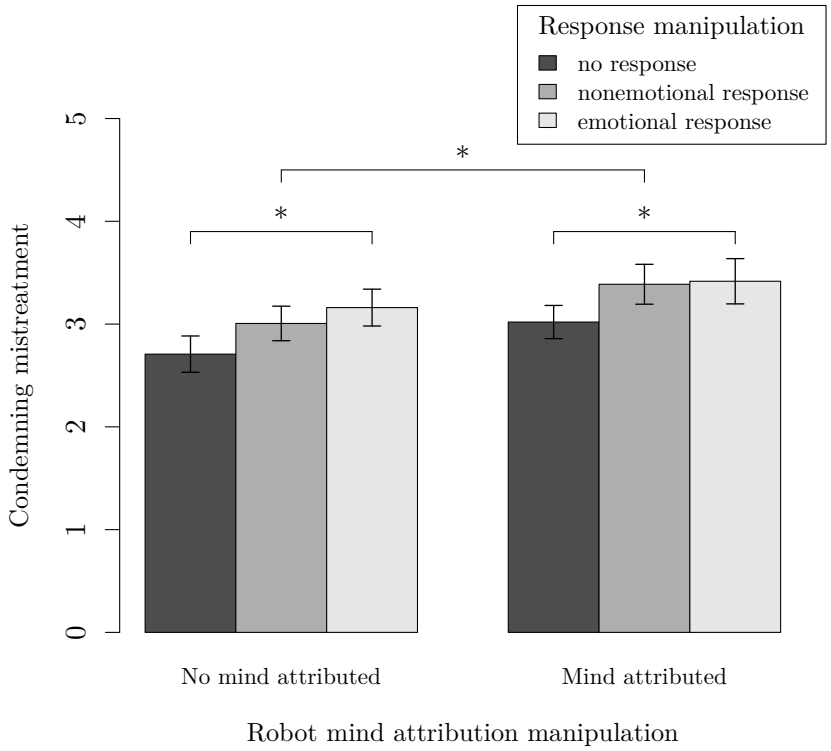


Figure 3.1: Difference in mean unacceptability for the robot introduction and robot response conditions. Contrasts marked with an * are significant at the $p < .05$ level in the post-hoc analyses.

5.68, $p = .004$). A correlation between age and unacceptability ratings however was not significant, $\rho = -.12$, $p = .175$). Thus, the difference in age between the conditions was not considered problematic. See Table 3.2 for the mean age per condition.

A Chi-Square test on the distribution of gender across the six conditions indicated that there was a significant difference in male to female ratio between the conditions, $\chi^2(5) = 13.32$, $p = .021$. However, a regression with unacceptability as dependent variable and gender as predictor indicated that gender was not significantly related to unacceptability, $t(124) = 1.83$, $p = .070$. Thus, the different gender ratios across the conditions were not considered problematic either. See Table 3.2 for the gender ratio per condition.

Individual tendencies to anthropomorphism were similar between the conditions, Fs < 1.79, ps $> .17$. Affinity with technology, as well, was similar between the conditions, Fs < 1.44, ps $> .24$. See Table 3.2 for the mean scores for each scale per condition.

Manipulation check A 2×3 ANOVA with robot introduction and robot response as independent variables, and explicit mind attribution as dependent variable indicated that the robot introduction had successfully manipulated mind attribution, $F(1, 187) = 12.46$, $p < .001$. There was no main effect for the robot response manipulation ($F(2, 187) = .73$, $p = .483$), nor a significant interaction effect ($F(2, 187) = 1.16$, $p = .315$). The mind attribution manipulation was thus deemed successful.

Main analyses

To test the influence of the robot's introduction on acceptability of robot bullying, a 2 (introduction: robot possessing mind vs not possessing mind)×3 (robot response: no response, non-emotional, emotional) ANOVA with 'unacceptability' as dependent variable was conducted. Significant main effects were found for both robot introduction ($F(1, 187) = 4.56$, $p = .034$) and the robot response ($F(2, 187) = 3.07$, $p = .049$). Since there was homogeneity of variance, post-hoc analyses with a Tukey correction were carried out.

Post-hoc analysis on the robot introduction manipulation revealed that participants in the robot possessing a mind condition reported finding robot bullying significantly less acceptable, M(*SD*) = 3.27(*.106*), than participants who had read the introduction where the robot was described as not possessing a mind, M(*SD*) = 2.96(*.105*); $t(1, 187) = -2.12$, $p = .035$. Post-hoc analyses of the robot response manipulation revealed a marginally significant difference between participants in the non-responsive condition, M(*SD*) = 2.86(*.126*), and the emotional response condition, M(*SD*) = 3.29(*.134*), $t(1, 187) = -2.32$, $p = .056$. The other contrasts were not significant, ts > -1.86, ps > .155. See Figure 3.1.

3.2.3 Discussion

Using an online scenario-based approach study, Experiment II.A explored the influence of a robot's mind attribution on how unacceptable people found it being bullied. Mind attribution was manipulated both directly, by telling participants that the robot was capable of thinking and feeling; and indirectly through having the robot respond to bullying in a way that implied mind. In line with hypothesis 1a, participants found bullying less acceptable if they had been explicitly told that the robot possessed a mind. In addition, and in line with hypothesis 1b, participants found robot bullying less acceptable when the robot responded in a negative, emotional way to the bullying. There was no interaction effect between robot introduction and type of response to bullying on how unacceptable participants found robot bullying.

These results suggest that perceived acceptability of robot bullying can be manipulated in two distinct ways. Moreover, the results indicate that even when people are explicitly told that a robot does not possess a mind, robot bullying is still seen as less acceptable when the robot responds with negative emotions. This suggests that explicit information on and implied robot mind affect acceptability in two separate manners. This would fall in line with the theory of empathy as a cognitive dynamic (Urquiza-Haas & Kotrschal, 2015). The implied robot mind (by an emotional response) would determine the initial evaluation of the bullying, and any explicit information on the robot's mind would subsequently adjust this evaluation. When the robot gave no response at all, the initial evaluation of how unacceptable the bullying behaviour was might return as "relatively acceptable"; but the final decision can still be adjusted by providing the participant with explicit information about the robot's mind.

A limitation of Experiment II.A concerns the scenario-based approach and the measurement of behavioural intentions rather than behaviour. Because Experiment II.A did not include actual human-robot interaction, behavioural intentions can only serve as a

proxy for participants' behaviour towards a robot. Previous research on the Media Equation theory (Reeves & Nass, 1996) demonstrated a divergence between self-report and actual behaviour. For example, research by Nass et al. (1994) showed that participants adhered to social norms of politeness when interacting with a computer. However, when asked if they ever were considerate of the computer's feelings, participants would strongly deny this.

In addition, Experiment II.A only measured whether robot mind attribution influenced how unacceptable robot bullying was considered. Whether considering robot bulling unacceptable leads to a reduction in abuse remains to be tested. We conducted Experiment II.B to overcome the problems surrounding scenario-based approaches, to replicate the findings, and to extend the experiment with a measurement of bullying behaviour.

3.3 Pilots

Experiment II.B followed a simple single-factor design with two dependent variables. The independent variable was mind attribution to the robot. The first dependent variable was how unacceptable robot abuse was considered, measured in the same way as in Experiment II.A. The second dependent variable was robot bullying, operationalised as participant's willingness to publicly humiliate the robot.

The mind attribution manipulation and robot bullying measure of Experiment II.B were validated on beforehand, by the means of two pilot studies.

3.3.1 Pilot 1: Robot mind attribution manipulation

The robot mind attribution manipulation of Experiment II.B was validated in Pilot 1. The mind attribution manipulation was an extension of the descriptions displayed at Table 3.1. A few lines were added about the robot's capabilities, as well as a picture of the Vector robot that would be used in Experiment II.B. See Table 3.3.

Method

Participants Participants were recruited via MTurk. 51 people participated. 24 participants read an introduction that depicted the robot as not possessing a mind, whereas 27 participants read an introduction that depicted the robot as possessing a mind. 54.90% of the participants were male, 41.18% female, and 3.92% (two participants) withheld from disclosing gender. Mean age was 39.12($SD = 9.07$). Participants received .65 US$ for their participation.

Procedure After assessing informed consent as well as age and gender, participants read either of the two proposed robot introductions and filled out the mind attribution questionnaire. Then they were thanked for their time and reimbursed.

Materials The mind attribution questionnaire was taken from Gray et al. (2007) and adapted so that the questions explicitly referred to the robot from the introduction. The

Table 3.3: Robot mind attribution manipulation as tested in Pilot 1 and subsequently used in Experiment II.B

Low mind attribution condition	High mind attribution condition
This is Vector, one of the robots that stay and work with us in the university lab. Being a part of the UC and our lab, he has appeared in a few of our demos before, although he cannot remember this. This is his first experiment, however. Vector is not aware that he is taking part in this, and as a robot he can't get excited or nervous about it. As a companion robot, Vector is programmed to appear to be social. His behaviour is coded to mimic an awareness of his surroundings as well as a personality, even though he possesses neither. For example, you can instruct him to store your name with a picture of your face, so that when he registers your face in a future encounter, he can retrieve your name and pretend to "recognise" you. There are a few other commands that have been pre-programmed, like to "go play" and "take a picture". However, Vector cannot understand or learn any new commands. Vector is also programmed to have different responses depending on how he is handled. For example, if he is handled carelessly, he will pretend to be upset and angry. If you stroke his back, he will start displaying signals of happiness. All these behaviours make Vector appear conscious. However, they are just animations: Vector is not capable of feeling any more than he is of telling whether he is being treated right or wrong on a moral level. Vector is outfitted with sensors that help him navigating around and create an illusion of spatial awareness to the observer. For example, these sensors force him to stop when he is about to bump into an object, and to slowly back away when a cliff is detected right in front of him.	This is Vector, one of the robots that stay and work with us in the university lab. Being a part of the UC and our lab, he is proud to have starred in a few of our demos before. This is his first experiment, however. Vector is very excited to be taking part in it, and also a little nervous. Vector is a social robot with his own personality: he is capable of processing, interpreting and emotionally responding to his environment. He can remember names and faces, and will recognise people he met before. He can understand different commands, like "go play" and "take a picture!". Also, his mood changes depending on how he is treated. For example, if he feels like he is treated unfairly, he will get upset or angry. Similarly, if you stroke his back, he will happily close his eyes and enjoy the back rub. He communicates his mental states through behaviour and his eyes. Moreover, Vector is aware of his surroundings. He sees what's in front of him, so that he avoids bumping into objects and will avoid driving off cliffs.

questionnaire measures to what extent the robot is capable of experiencing 18 different emotional and cognitive states, using a 5-point Likert scale that ranges from "very incapable" to "very capable". See Appendix A for the questionnaire.

Results

Mind attribution to the robot was significantly higher for the introduction depicting the robot as possessing a mind, M(SD) = 2.64(*.91*), than for the introduction depicting the robot as not possessing a mind, M(SD) = 1.90(*.95*), $t(48.75)$ = -2.85, p = .006. The mind attribution manipulation was thus considered valid to use.

3.3.2 Pilot 2: Robot bullying

Experiment II.B operationalised its dependent variable "robot bullying" as whether participants chose to humiliate the robot they just interacted with by picking a condescending review to be put up on display next to it. This operationalisation was developed, tested, and validated in Pilot 2. The resulting review pairs were similar in sentiment and informativeness, but differed significantly in how condescending they were towards the robot.

Method

Participants The reviews were constructed in two rounds of pilots, then validated in the third. All participants were recruited via MTurk. For the third and final round of piloting the reviews, 45 participants were recruited. Two participants were excluded because of straightlining, i.e. answering every item on the survey with the same score; thus resulting in a dataset of 43 participants. 65.11% of the participants were male, 30.23% were female, and 4.65% withheld from disclosing their gender. Mean age was 41.17($SD = 10.43$) years. Participants received .60 US$ for their contribution.

Procedure For the first round of testing, 12 reviews were constructed out of actual reviews of the robot, taken from different websites. These reviews were rated on each of three scales: the sentiment expressed (ranging from "very negative" to "very positive"); how informative the reviews were for someone contemplating purchasing a Vector robot (ranging from "not informative at all" to "very informative"); and finally, how condescending each of the reviews was (ranging from "very condescending" to "not condescending at all"). 5-point Likert scales were used to collect participants' responses.

After the first stage of testing, five pairs of reviews were selected which were rated roughly equally high with regard to affect and usefulness, but diverged in how condescending they were to the robot. Those reviews were adjusted to further decrease any differences in affect and usefulness scores, and retested. One of the pairs was dropped as the difference in condescension ratings was only marginally significant, resulting in a final set of four pairs of reviews (two positive, two negative) that were equally positive/negative and informative, but significantly different in how condescending they were of the robot. This set was then tested a third and final time.

Results of the third round of testing

Four pairs of reviews of the robot (two positive and two negative) were tested by the means of t-tests on being equal in sentiment expressed and informativeness, but different in how condescending they were towards the robot.

Three of the four review pairs were similar on sentiment expressed, $-1.92 < t\text{s} < -0.18$, $p\text{s} > .062$. Of these three pairs, two were seen as equally useful, $-1.49 < t\text{s} < -.57$, $p\text{s} > .147$. These two pairs differed significantly in how condescending they were perceived to be, $t\text{s} > 3.28$, $p\text{s} < .002$. Both pairs were positive in overall sentiment.

Since these were the only reviews that differed exclusively on how condescending they

were, the negative reviews were disregarded as a measure of humiliation. The two positive review pairs were considered valid to use.

3.3.3 Conclusion

In two pilot studies, the experimental manipulation and one of the two dependent measures for Experiment II.B were validated. Pilot 1 confirmed that providing people with an introduction that depicted the robot as capable of thinking and feeling increased their subsequent mind attribution to that robot, compared to people who read an introduction that explicitly stated the robot did not possess the ability to think and feel. This manipulation was thus adopted for Experiment II.B.

Pilot 2 developed and validated the operationalisation of the robot bullying measure. The objective was to find four pairs of reviews, where within each pair, both reviews were equally positive or negative in sentiment expressed, and equally informative on the robot. The only difference would be how condescending in tone those two reviews were towards the robot. The underlying rationale was that if participants in Experiment II.B would choose the condescending review over the equally useful alternative to be displayed next to the robot, this could be taken as an attempt to humiliate the robot.

After two initial rounds of testing and revision, two pairs of positive reviews (so four reviews in total) were validated. These two review pairs were thus used as a measure of robot bullying in Experiment II.B.

3.4 Experiment II.B

Experiment II.B was designed to include a human-robot interaction with Vector, a social robot (see Figure 3.2). There was one manipulation, i.e. robot mind attribution in the introduction, and two dependent variables, i.e. unacceptability of bullying the robot and whether or not participants chose to publicly humiliate the robot.

3.4.1 Method

Participants

Participants for Experiment II.B were recruited on campus, through posters, online recruitment, and snowballing. 67 people participated. 41.79% of the participants were male; 57.72% were female, and 1.49% (one participant) did not identify as either gender. The average age was 25.46 (SD = 6.18) years. In return for their participation, participants could enter a draw to win a 50$ gift card for a local shopping mall. In addition, a bowl of candy from which the participants could take freely was offered during the trials.

Materials

Vector Vector is a social companion robot, produced by the consumer robotics company Anki. Vector is a small (9×6×7 cm) robot which comes pre-programmed with fully autonomous behaviour. It responds to specific commands as well as physical interaction,

Figure 3.2: Left; the Vector robot. Right; the study set-up.

and will move around and "explore". Vector does not talk under most circumstances, although it can be taught names and will respond verbally to the command "what's the weather?". In general, however, Vector communicates through behaviour, facial expression and a variety of chirping and buzzing noises. Its behaviour is animated and although it does not have a mouth, it can create a range of facial expressions with its eyes. Vector comes with an Anki cube, which it can interact with (e.g. picking it up). See Figure 3.2.

Measurements

Individual differences in anthropomorphism tendencies and affinity with technology Like in Experiment II.A, trait anthropomorphism (Waytz, Cacioppo, & Epley, 2010) and affinity with technology (Neyer et al., 2012) were assessed to check whether participants between the two conditions were similar. See Appendix A for the full questionnaires.

Unacceptability of robot bullying During the interaction with Vector, participants were asked to try out a total of three commands and behaviours that would elicit a negative response from Vector. The unacceptability scale consisted of seven items. Six referred directly to the negative commands and behaviour participants had been asked to try on Vector, asking the participant "How acceptable would it be if someone were to carry out [*behaviour or command*] repeatedly for an extended period of time?" and "How important is it to protect Vector from being [*behaviour or command*] repeatedly for an extended period of time?". The final question asked how important it was that Vector was treated humanely in general. Responses were measured on a 5-point Likert scale ranging from "completely unacceptable"/"very unimportant" to "completely acceptable"/"very important". As in Experiment II.A, the scale was scored in such a way that a high score corresponded with participants finding the behaviour less acceptable.

Humiliating Vector After the interaction with Vector, participants were asked to select one review out of four, which – according to the cover story – would be put up in public next to Vector on the upcoming Open Day of the lab. The four reviews in fact consisted

of two review pairs which were equally informative and positive, but different in how condescending they were towards Vector (see Section 3.3.2).

Behaviour interpretation check During the interaction with Vector, participants were asked to try out a total of three commands and behaviours that would elicit a negative response from Vector. These commands were telling Vector he was a bad robot; holding Vector up in the air so its wheels could not touch a surface; and picking Vector up, turning it upside down, and shaking it violently.

To check if participants indeed interpreted the responses to each of these three behaviours as negative, nine questions were included in the survey. For each of the three behaviours, participants indicated how positive or negative Vector responded, how the response made the participant feel, and how willing they would be to repeat the behaviour/command a number of times in a row. None of the participants (incorrectly) interpreted the behaviours as positive.

Procedure

Participants were seated at a table with a Vector robot asleep on its charger, an Anki cube, a folder with the information sheet and consent form, and a tablet. See Figure 3.2 for the study set-up. The experimenter gave a brief introduction: thanking participants for their participation; clarifying any issues that participants might have after reading the information sheet; demonstrating how to give Vector a voice command; and explaining that the tablet would take the participant through the procedure step by step. After ensuring that the participant was ready, the experimenter left the room.

First, participants were instructed to report their demographics (i.e. age and gender). Then, participants were randomly assigned to one of the two introductions to the Vector robot (manipulation: high or low mind attribution). Subsequently, participants were given a list of voice commands and behaviour to practice with Vector. Upon the first command, Vector would wake up and drive from its charging dock onto the table. Some of the commands and behaviour evoked a negative response from Vector (e.g. telling it "Bad robot!" or lifting it in the air), some evoked a positive response (e.g. telling it "give me a fist bump!" or stroking its back). Being an autonomous robot, Vector was animated throughout the interaction. When it had not received a command, it would appear to be entertaining itself, roaming around, sometimes looking up to the participant and make giggling noises, or picking up its Anki cube and dropping it off at another spot.

After 10 minutes of interaction with the Vector robot, the list of commands on the tablet disappeared and was replaced by an instruction for the participant to put Vector back on its charging dock and continue with the survey part of the experiment. The survey assessed (in order) anthropomorphism (Waytz, Cacioppo, & Epley, 2010), affinity with technology (translated from Neyer et al., 2012), the set of reviews, the control questions on Vector's behaviour interpretation, and the unacceptability questions.

At the end, participants were instructed to call the experimenter in again. The experimenter thanked them for their time, verbally debriefed them, and gave them a raffle

Table 3.4: Mean(*SD*) per condition for age, percentage female, and each of the different scales

	Low mind attribution	High mind attribution
Age	25.24(*6.83*)	25.68(*5.56*)
Percentage female	45.46%	67.65%
Trait anthropomorphism	4.19(*1.32*)	4.54(*1.31*)
Affinity with tech	3.72(*.89*)	3.77(*.59*)
Condemning mistreatment	3.57(*.95*)	3.98(*.62*)

ticket for the 50$ voucher draw. The entire experiment took between 20 and 30 minutes.

3.4.2 Results

Preliminary analyses

To check the internal consistency of the scales used in Experiment II.B, we computed Cronbach's alpha. Alphas ranged from acceptable to good (Cronbach, 1951): for individual differences in anthropomorphism $\alpha = .76$, affinity with technology $\alpha = .75$, behaviour interpretation $\alpha = .77$ and unacceptability of bullying $\alpha = .88$.

Four t-tests were carried out to ensure that participants between the conditions had interpreted Vector's behaviour in the same way, were equally inclined to anthropomorphise, did not differ in their affinity with technology, and were of similar age. No significant differences were found: $t(64.05) = .18$ and $p = .854$; $t(64.92) = -1.11$ and $p = .272$; $t(55.19) = -.26$ and $p = .794$; and $t(62.94) = -.40$ and $p = .692$, respectively. A Chi-Square test indicated that the distribution of males and females was equal between the high and low mind attribution condition, $\chi^2(2) = 3.96$, $p = .14$. Randomisation was thus considered successful. See Table 3.4 for the descriptives.

Levene's test was significant, $F(1, 65) = 4.15$, $p = .046$, indicating that the variances were not equal between the high and low mind attribution condition. As a result, Welch approximation of degrees of freedom was used for the main t-test.

Main analyses

To test whether the robot's mind attribution manipulation had an effect on how acceptable participants found robot bullying, an independent samples t-test was conducted. Participants in the high mind attribution condition found bullying Vector less acceptable, M(*SD*) = 3.98(*.62*), than participants in the low mind attribution condition, M(*SD*) = 3.57(*.95*). This difference was significant, $t(54.88) = -2.09$, $p = .042$.

To test for a difference in selecting condescending reviews to be put up in public next to Vector, a logistic regression was run with 'picked a condescending review' as a dichotomous dependent variable and condition as a predictor. Participants in the low mind attribution condition were equally likely to put up a condescending review, $z = 1.12$ $p = .262$. Condescending reviews were no more common for participants in the low mind

attribution condition (12 out of 33) than in the high mind attribution robot condition (17 out of 34).

Exploratory analyses

Two exploratory analyses were conducted. The first tested if attributing the robot a mind had led participants to prefer any of the four reviews. The second tested whether there was a direct relationship between participants' unacceptability scores of robot bullying and selecting a condescending review.

Firstly, a Chi-Square test was conducted on the distribution of selected reviews between conditions. No difference was found, $\chi^2(3) = 1.59$, $p = .661$, indicating that the mind attribution manipulation had not affected participants' review selection.

Secondly, a logistic regression with 'condescending review' as a dichotomous dependent variable and 'rating robot bullying unacceptable' as a continuous predictor was performed to test if there was a relationship between finding robot bullying unacceptable and selecting a humiliating review. There was a positive relationship between selecting a condescending review and finding robot bullying unacceptable, $z = 2.75$, $p = .006$. People who had selected a condescending review also found bullying less acceptable.

3.4.3 Discussion

In Experiment II.B, participants interacted with a Vector robot which was introduced as either high or low in mind attribution. There were two outcomes of interest: how unacceptable participants rated bullying Vector post-interaction, and whether participants elected to publicly humiliate Vector.

Hypothesis 2a was confirmed: the findings from Experiment II.A were replicated. When the robot was attributed a mind, participants considered bullying it less acceptable. However, hypothesis 2b was rejected. Intriguingly, in spite of this the mind attribution manipulation did not influence participants' choice in putting up a condescending review next to Vector (even while provided with a less condescending alternative).

Unplanned exploratory analyses found a relationship between condemning bullying and selecting a condescending review, with the chance of a participant selecting at least one condescending review to put up increasing as they found bullying less acceptable.

3.5 Main discussion

In Experiments II, the relationship between explicit and implied robot mind attribution and perceived acceptability of robot bullying was investigated. In two experiments, a robot was introduced as either high or low in mind attribution. Acceptability of robot bullying was measured in both experiments. In Experiment II.B participants were also offered the option to publicly humiliate the robot.

In line with our predictions, a higher mind attribution to the robot, manipulated either by explicitly informing participants that the robot was capable of thinking and feeling or by having the robot respond in a negative emotional way to bullying, led to lower

acceptance of robot bullying. Even when participants were told that the robot did not possess a mind, robot bullying was still found less acceptable when the robot responded in a negative emotional way.

These findings are in line with the theory of empathy as a cognitive dynamic (Urquiza-Haas & Kotrschal, 2015), which states that empathy is the result of an initial implicit evaluation, that is subsequently adjusted by a deliberate and reflective process. The robot's response to bullying would set a preliminary empathetic response, which would then be further nuanced based on the participant's cognitive appraisal of the robot's mind attribution. Moreover, these results tie in with previous findings from mind perception (Gray et al., 2007), and dehumanisation theory (Haslam, Loughnan, et al., 2008), both of which associated mind attribution with the subject being more deserving of moral treatment.

However, there are two alternative explanations for the findings from Experiment II.A: Firstly, maybe people consider all kinds of robot bullying bad, or maybe it was the abuser disregarding the robot's request for them to stop that made the action immoral. However, if either of those had been the case, robot bullying in the scenarios where the robot did not respond or responded in a non-emotional way would have been equally unacceptable as bullying in the emotional response scenario.

Secondly, participants may have been unsure about whether the alleged abuser knew whether the robot possessed a mind, and found the bullying unacceptable because the abuser behaved in a hurtful way to an agent that to his best knowledge was capable of feeling. But then one would have expected no difference between the high and low mind attribution condition in robot bullying acceptability ratings when the robot appeared to possess a mind (i.e. when it gave an emotional response). No such interaction effect was found.

A second interesting finding was that the mind attribution of the robot did not affect whether people chose to publicly humiliate it. Exploratory analyses showed a positive relationship between finding robot bullying unacceptable and belittling it in public. This is intriguing, as common sense would suggest this relationship to be inverted: if people perceive an agent to be of higher moral standing, they should be less likely to abuse it. On this note, it is interesting to note that Tan et al. (2018) found a marginal trend where bystanders of robot bullying were less likely to intervene when the robot did (versus did not) give off emotional cues. This was in spite of them rating the robot as more capable of experiencing human emotion than its non-emotional version. The paper does not report on a relationship between acceptability of robot bullying and intervention tendencies, but our present research suggests that emotional cues would have decreased how acceptable participants found the bullying.

At the same time, a certain level of robot "cuteness" might have influenced both its right to be protected and people's tendency to belittle it. Mind attribution has been divided in the Human Nature (or Experience; essentially the capability to feel) and Uniquely Human (or Agency; roughly said the capability to think) dimensions before (Gray et al., 2007; Haslam & Loughnan, 2014). While Gray et al. (2007) related the former to the

right to be protected and the latter to the duty of being held accountable for one's own responses, Haslam and Loughnan (2014) linked both to moral standing. If we take the middle road and assume that both perceived Human Nature and (to a lesser extent) Uniquely Human traits in the robot enhanced its moral standing, but only Human Nature increased people's tendency to belittle (or baby-talk) it, then that could be an explanation why manipulated mind attribution increased moral standing but not people's tendency to humiliate the robot.

The results from Experiment II shed new light on the complicated relationship between robot mind attribution, perceived right of protection from abuse, and willingness to belittle the robot. As shown in previous research (for example Bartneck, Verbunt, et al., 2007; Horstmann et al., 2018; Tan et al., 2018), empathy with a robot may not necessarily lead to a lower willingness to harm the robot. The current experiments suggest that this may be because moral acceptability and bullying behaviour are not related, or at least not in the way one would expect. This has great implications for researchers in the field of HRI, who may take empathy as an operationalisation of prosocial behaviour tendencies. It also opens up a whole new venue of potential research on how people mitigate morality and mind perception in the face of robot bullying. Experiments IV and V will explore the relationship between robot mind perception and robot bullying further.

3.5.1 Limitations

The measure of robot bullying used in Experiment II.B was rather subtle, as participants chose between four generally positive reviews. The measure had been inspired by the measure of derogation in Dahl, Vescio, and Weaver (2015). In this study, participants had been asked to select a more or less objectifying avatar to represent their online female teammate. Choosing a sexualised depiction of the female avatar was interpreted as a measure of aggression and condescension. In the current research, instead of choosing between depictions with varying levels of objectification to be put up next to a teammate, participants chose between reviews with varying levels of condescension to be put up next to the robot. However, making a depiction more or less exposing can be done quite simply. Creating reviews that differ in condescension but are otherwise identical, on the other hand, is less straightforward (as also indicated by the three rounds of pilots that were needed). Pilot testing increased confidence in a successful manipulation of humiliation levels, but the question remains if participants who selected the more humiliating review with the intention of humiliating the robot – which is essential to the definition of bullying. In addition, the measure did not incorporate the component of repetition which is generally considered a characteristic of bullying (see Section 1.2). Future studies should thus explore alternative ways of operationalising robot bullying. For example, alternative ways of humiliation could be devised, an element of repetition should be added, or scholars could more directly address verbal and physical abuse.

Due to human error, data on age and gender was not collected in Experiment II.A for the first 66 participants. In addition, randomisation of age and gender failed in Experiment II.A. The failed randomisation was considered not problematic as neither age nor gender

was significantly related to acceptability of robot bullying, the sole dependent variable of Experiment II.A. An alternative solution to the failed randomisation would have been to include age and gender in the further statistical analyses as covariates. However, due to the partially failed data collection on age and gender this approach would have severely compromised statistical power.

3.5.2 Conclusion

People bullying autonomous robots in public is a surprisingly common phenomenon but so far there is little understanding of the psychological motivation to this behaviour. In Experiment II we manipulated robot mind attribution and measured its effect on how acceptable robot bullying was deemed as well as how willing people were to bully the robot themselves. The results showed that while robot mind attribution influences how acceptable people find bullying it, it does not make them more or less likely to bully the robot themselves.

The findings imply that enhancing feelings of empathy with a robot may not necessarily make people less prone of abusing it. These findings are highly relevant for the development of autonomous robots in a social setting. Such robots will likely need to be designed with different strategies of how to discourage robot bullying in mind, and in spite of what common sense may suggest, making a robot appear to possess a mind does not seem to discourage robot bullying. In addition, the results are relevant for HRI researchers focusing on robot likeability and user behaviour. Measurements of moral acceptability of robot bullying may not be a valid predictor of actual bullying behaviour.

Chapter 4

Study III: The Cleverbot Studies

An adapted version of this chapter has been submitted for review with Interaction Studies, under the title 'Correlates between chatbot humanlikeness and abuse'.

4.1 Introduction

Experiments I and II (Chapter 2 and 3, respectively) reported on empirical experiments. Study III concerns observational data. While the lack of experimental design restricts the inferences that can be made, the observational setup does provide the chance to cross-reference the findings from the previous studies. Specifically, Study III will test for correlations between a chatbot's humanlikeness and user abuse.

As discussed in greater extent in section 7.2, studying robot bullying in a controlled experiment has the drawback that participants, quite understandably, often feel observed and self-conscious of their behaviour. As a result, participants are likely to adjust their behaviour; either so that it is more exemplary than it normally is, or in such a way that they behave in the way they think the experimenter wants them to. In other words, participants display socially desirable behaviour, which may bias the results (Nederhof, 1985; Ritter & Eslea, 2005).

Especially when the measure of interest is something as objectionable as bullying, social desirability will interfere with the participants' inclination to display any such behaviour. In controlled experiments bullying therefore is normally operationalised as a subtle form of aggression. While experimental designs are central to empirical research, their value depends on how well this operationalization for bullying behaviour generalises to actual bullying scenarios. Therefore, observational data can be useful to provide findings from controlled experiments with external validity.

Previous studies on human-chatbot interactions found that as many as one in ten conversations contain the user threatening, assaulting, or otherwise bullying the chatbot (De Angeli & Brahnam, 2008; De Angeli & Carpenter, 2005). Users engage in chatbot abuse with minimal or no provocation, and in spite of knowing that the agent they're interacting with is a non-sentient AI and that as of such abuse cannot affect its mental

With tremendous thanks to Rollo Carpenter, who graciously shared conversation logs as well as his insights from owning and moderating public chatbots for over two decades for this experiment.

state. As chatbots get increasingly common in daily interaction — for example AI personal assistants like Siri and Cortana as well as basic customer service with major companies — ethical and practical concerns can be raised about humans verbally abusing chatbots (Brahnam, 2005; De Angeli, 2009).

To date, chatbot abuse has been described by various scholars (see for example De Angeli, 2009; Strait, Contreras, & Vela, 2018) but few empirical studies have been published on what motivates users to engage in abusive behaviour (but see Brahnam & De Angeli, 2012; De Angeli & Brahnam, 2006; De Angeli, Johnson, & Coventry, 2001). Study III will therefore investigate whether there is a correlation between specific characteristics of the conversation between a human and a chatbot, and human verbal aggression towards that chatbot. More specifically, we will test the relationship between Cleverbot's ability to maintain a facade of humanlikeness and specific instances of user abuse.

4.1.1 Literature

Incidence: nonhuman conversation agents suffer more abuse than humans

Users can get surprisingly obscene when interacting with chatbots (see for example De Angeli & Brahnam, 2008; De Angeli & Carpenter, 2005; Hill, Ford, & Farreras, 2015; Strait et al., 2018). In the 100 conversations that Hill et al. (2015) analysed, 4.29% of the messages directed towards the chatbot contained profanity of some sort. Moreover, this statistic was not the result of a small number of exceptionally abusive users: 80% of the conversations contained profanity. In comparison, 15% of the human-to-human conversations that were analysed in the same study contained some sort of foul language (Hill et al., 2015).

Perhaps due to Hill et al. (2015)'s definition of profanity (they counted all words that would be rated as PG or above), other studies on the topic of chatbot abuse have found slightly lower incidence rates for chatbot abuse. Brahnam and De Angeli (2012) analysed anonymous conversations between users and an internet chatbot and detected profanity in "only" 54% of all user-chatbot conversations in their database, and sexual references in about 65%. Veletsianos, Scharber, and Doering (2008) analysed (non-anonymous) student-chatbot interaction in an educational setting, where the chatbot was supposed to help tutor middle school students. Around 40% of all comments that students made towards the chatbot tutor were "unacceptable", with roughly 45% of all unacceptable comments being sexually explicit.

Lortie and Guitton (2011) studied conversations from the Loebner competition, an annual contest between computer programs on which appears the most "human" in a conversation (Mauldin, 1994). All conversations analysed in Lortie and Guitton (2011) were between humans; yet the human judges were more aggressive in their conversation when they thought they were interacting with a robot. This phenomenon occurred despite the fact that their interaction partner did not display any verbal aggression (Lortie & Guitton, 2011).

Ethical implications

At the current state of technical development, chatbots are insentient and verbal abuse directed towards them will not affect them. Thus, as far as the well-being of the chatbot is concerned, verbal abuse of chatbots is not problematic long as the verbal aggression takes place in private and does not cause offence to any onlookers. Yet there are some ethical implications of chatbot abuse have been raised in the literature.

Some people would already disagree with the assessment that any behaviour is acceptable as long as it occurs out of sight and does not get anyone hurt (Darling, 2012; Whitby, 2008). In addition, (Brahnam, 2005) argued that accepting inappropriate and offensive language towards conversational agents could encourage users to abuse human interaction partners in a similar setting. De Angeli and Brahnam (2006) found that presenting a chatbot as male or female evoked user responses that were in line with gender stereotypes. They argued that if users shape their behaviour to a bot based on their experience with fellow human interaction partners, it seems plausible that humans orient their behaviour towards other humans based on what behaviour is deemed acceptable towards robots and other nonhuman entities. Strait et al. (2018) indeed found that verbal aggression towards robots correlated with overall aggressive tweeting behaviour on Twitter, suggesting that abuse of robots and abuse of humans are related. However, as there was no experimental manipulation involved, causal inferences on whether verbal abuse of robots could generalise to abuse of humans cannot be made.

On a surface level, this argument appears similar to the assertion that violent video games cause violent behaviour. In spite having been hotly debated over the last decades, no consensus has been reached yet on whether this relationship actually exists (compare for example Anderson & Bushman, 2001; Anderson et al., 2010; Ferguson, 2007). However, in video games the violence is a means to an end: the monster has to be slain, the enemy defeated, points or loot collected. Aggression in and of itself is not related to players' motivation to game; rather, aggressive scenarios such as fights or wars provide an easy template for the set of tasks, puzzles, and challenges that constitutes the game experience (Przybylski, Rigby, & Ryan, 2010). Abuse of agents, on the other hand, does not have a higher goal. There is no gain to be had from insulting and sexually harassing AI's other than the enjoyment of the behaviour itself.

Moreover, these chatbots are a representation of a social agent not just to the abuser, but also to others. Abuse of AIs, as well as the acceptance of abuse by AIs could thus be considered offensive to third parties. While those people would not sit in on abusive interactions themselves, a lack of appropriate response from the chatbot to abuse could spark significant resentment. This has in fact already happened in the context of sexual abuse of AI personal assistants. In 2017, journalists tested the responses of female personal assistants and reported how all AI assistants tested would either not understand sexual abuse, or jokingly brush it aside (Curry & Rieser, 2018; Fessler, 2017b). This reveal was especially painful as it happened at the height of the #MeToo debate. Less than a year later, the tech companies behind those AIs had received major push back from their customers, who demanded that more appropriate responses would be installed on

the AIs (Fessler, 2017a), and at least one company adjusted their algorithm to respond more sternly to sexual harassment (Hern, 2010).

In addition to the ethical issues — whether some behaviours are deemed immoral even if they happen in private, if abusing an agent that represents a human could lead to abuse of humans, and if tolerating abuse towards an agent that represents a group would offend human members of that group — that have been raised, Brahnam (2005) also points out that chatbot abuse may have practical implications when the chatbot is part of a company, e.g. as the first contact point in customer service. Abusive communications between a user and a customer service chatbot could damage the relation between the company and the customer, as the customer may henceforth associate the company with the negative interaction. While this may not be a problem for the abuser, it may damage the company's reputation.

Need for a theoretical framework of chatbot abuse

As it has been estimated that by 2020 20% of customer service operations will use virtual agents (Moore, 2018), understanding and developing effective strategies to deal with chatbot abuse by users is getting increasingly relevant. While the greater incidence of abuse, sexual harassment, and overall profanity in human-chatbot interaction has been widely noted and raised ethical concerns, only few researchers have studied where the users' motivation to engage in such behaviour originates from.

In De Angeli et al. (2001), ten participants interacted with chatbot Alice over the course of a week and discussed their experiences in a focus group. The authors noticed that participants would often insult the chatbot during chat sessions. In the focus group participants explained that they had wanted to secure an asymmetrical relationship with the chatbot where they were dominant and Alice was submissive, and that verbal abuse helped them create a position of power over Alice. Why the participants desired to be in such a position (or why they felt verbal aggression was the way to get there) was left undiscussed.

De Angeli and Brahnam (2006) studied the influence of gendered embodiment (male, female, or gender-ambiguous) of a chatbot on users' sexual abuse. When the chatbot was framed as female, as many as 18% of all conversations focused on sex, versus 2% for the male chatbots and none of the conversations for the asexual chatbots. Moreover, users would often make aggressive demands towards the female chatbot whereas the male embodiment got more questions about its sexual propensity. This was in spite of the chatbot not engaging in any sexual role play or discussion initiated by the users. While this study strongly suggested that anthropomorphic qualities of the robot were related to robot abuse, it did not answer the question whether these are the reason for the abuse or simply a prerequisite.

Brahnam and De Angeli (2012) explored theoretical explanations for chatbot abuse. They noted that verbal interaction with conversational agents appeared to provide an ideal environment for aggressive and sexual disinhibition, as computer-mediated communication reduces social pressure, thereby liberating individuals from any boundaries and constraints

imposed by face-to-face conversation. Moreover, being in a virtual environment reduces the victim's ability to retaliate and increases the ease with which the victim is dehumanised. Those factors, as well, would enhance disinhibition (the online disinhibition effect; see also Lowry et al., 2016; Suler, 2004). This is in line with the reasoning of Hill et al. (2015), who found a 30-fold increase in profanity when users went from interacting with someone familiar to someone anonymous. The authors suggested that the anonymity of the user had led to their greater proclivity for verbal abuse of the chatbot, although they also admitted that the increase in profanity was well beyond what they would have expected.

Establishing factors that discourage chatbot abuse may help provide insight in what motivates the behaviour in the first place. However, few studies have been dedicated to discouraging chatbot abuse. Chin and Yi (2019) invited participants to abuse a conversational agent and studied if different types of response to this abuse (ignoring it, displaying empathy, counter-attacking) influenced users' reported emotional state after the interaction. They found that empathetic responses to abuse evoked the highest levels of guilt and shame and the lowest levels of anger in participants. However, these users were explicitly instructed to mistreat the agent; there is no way of telling whether a similar pattern will occur for spontaneous abuse. In addition, no explanation was given for why users engage in abuse in the first place.

So far, gender of a chatbot (Brahnam & De Angeli, 2012; De Angeli & Brahnam, 2006) and dispositional aggression of the user (Strait et al., 2018) have been empirically studied. However, factors that are related to the conversation behaviour of the agent have been largely ignored. More specifically, to our knowledge no research so far has covered whether there is a correlation between the chatbot's behavioural humanlikeness and chatbot abuse. This is a relevant question, however, since on one hand there is the push for developing a chatbot that is truly indistinguishable from a human, while the work by De Angeli and Brahnam (2006); Lortie and Guitton (2011) suggests that this will not prevent abuse, and may even encourage it.

4.1.2 Current study

Research questions

Study III harvested and analysed the content of 283 conversations that took place between the Cleverbot online chatbot and any one anonymous user. The research questions were as follows:

1. Is there a relationship between humanlikeness of the chatbot, and verbal abuse by the user?

2. Is there a relationship between humanlikeness of the chatbot, and sexual abuse by the user?

Three measures were taken as influencing apparent humanlikeness in Cleverbot. The first measure was whether third party observers, who were left uninformed of the identity of the two conversationalists, could identify Cleverbot as a chatbot.

The second indication of humanlikeness of Cleverbot was the number of nonsensical responses it gave. In a dialogue, conversation partners tend to engage in a collaborative interaction. This means that when formulating a response, each conversation partner has to recognise the intention or goals that were expressed by their fellow partner, and formulate a response that takes these into account (Ardissono, Boella, & Lesmo, 2000). If one of the two conversationalists fails to acknowledge their partner's contribution, this reduces how humanlike they are seen (Lortie & Guitton, 2011). Thus, the more nonsensical responses Cleverbot would give, the less humanlike it would appear.

The third indication of humanlikeness of Cleverbot was the chatbot claiming to be human. While all chatbots aim to mimic human behaviour, actually claiming to be human is generally considered unethical and off-limits, to the point where it is a forbidden move in the Loebner competition (Lortie & Guitton, 2011). Cleverbot attempting to convince the user of its humanity by boldly stating to be a human could thus be seen as a strategy that only humans would use.

Alternatively, however, claims of humanity by Cleverbot could be perceived as a blatant attempt at deceit. This makes it hard to determine whether any effects of this measure would be due to an increase in humanlikeness, or to humans responding to an obvious lie. Thus, to control for this second possibility, Cleverbot claiming the human user to be a chatbot was also measured as a control variable. The idea behind this additional measure was that if user aggression is influenced by claims of humanity from Cleverbot *because* the user interprets these claims as a lie, there will also be a relationship between user aggression and another instance of Cleverbot lying, i.e. claims of Cleverbot that the user is a chatbot. In the absence of such a relationship, claims of humanity can be interpreted as contributing to humanlikeness.

In addition, the number of instances where Cleverbot insulted the user was measured as a predictor. Since the chatbot's lexicon does not include swear words, insults mostly take the form of Cleverbot accusing the user of lying. It seemed likely that this would provoke an aggressive response from some users.

Finally, a measure of self-disclosure by the user was included, as self-disclosure was considered an indication of trust and friendliness (Dindia, Fitzpatrick, & Kenny, 1997; A. Ho, Hancock, & Miner, 2018). This was operationalised whether the user indicated their gender to the chatbot.

Hypotheses

The hypotheses were as follows:

1. Based on the work by De Angeli and Brahnam (2006); Lortie and Guitton (2011) it was hypothesised that humanlikeness of the chatbot would be positively related to chatbot abuse.

    (a) A third party judging Cleverbot to be human, and number of claims by Cleverbot to be human were expected to be positively related to chatbot abuse.

(b) Number of nonsensical responses by Cleverbot was expected to be inversely related to chatbot abuse.

2. No relationship was expected to exist between Cleverbot claiming that the human was a chatbot and user abuse.

3. It was expected that insults from Cleverbot would be related to higher counts of chatbot abuse.

4. In line with research on the positive relation between self-disclosure and enjoyment of the conversation (A. Ho et al., 2018; Lortie & Guitton, 2011), it was expected that self-disclosure by the user would be related to lower instances of chatbot abuse.

4.2 Method

4.2.1 Procedure

Two stepwise regressions were performed with abuse of Cleverbot and sexual comments made towards Cleverbot as dependent variables. In each conversation, the instances of verbal aggression towards Cleverbot were counted; as well as the number of times Cleverbot made a nonsensical reply; the number of times Cleverbot insulted the user; the number of times Cleverbot claimed to be a human; the number of times Cleverbot claimed the user to be non-human; and how often the user made an explicit sexual remark. If the users mentioned their gender, this was also coded. For the approximate Loebner test, for each conversation up to four naive participants were recruited from Amazon Mechanical Turk and asked to indicate whether either of the people in the conversation had been a chatbot.

4.2.2 Dataset

Cleverbot

The chatbot that provided the data for Study III is Cleverbot[1], which is based on the award-winning Jabberwacky[2] engine. The chatbot Cleverbot does not make use of any state-based machines or scripts to carry out its conversation. Rather, it repeats things previous human users have said to it in a similar context. In combination with the chatbot's incapability to "remember" previous exchanges in the conversation, this "parroting" method of holding a conversation means that chats with Cleverbot often turn out extremely realistic or bizarre.

Data collection

Cleverbot users are informed on the homepage that the data they submit to the chatbot may be processed[3]. For Study III, an anonymous dataset was obtained from the owner of the Cleverbot website, Rollo Carpenter, upon request.

[1] `https://www.cleverbot.com`
[2] `http://www.jabberwacky.com`
[3] `https://www.cleverbot.com`

The conversations were collected between midnight and 00.30 UTC, and midday and 12.39 UTC on 18 April 2018 by Rollo Carpenter. These timings were chosen so that both users from the USA and from Europe would be included. Rollo Carpenter removed conversations that went on for fewer than 20 turns, conversations that were not conducted in English, and conversations that "launched straight into vulgar sexual language with no pretence at, or attempt at, conversation" [personal communication, 18.05.2018]. The resulting 283 conversations were provided for the first author for analysis in simple text format and without any identifying information (e.g. time stamp, location, IP address).

4.2.3 Coding procedure

Given the issues encountered with analysing the data through a sentiment analysis tool (see Keijsers, Kazmi, et al., 2019), coding was done by hand. All the conversations were coded by the first author. With the exception of gender, all variables were count variables. A second coder, not otherwise involved in the current research, re-coded 30 randomly selected conversations to establish inter-rater reliability. The dataset is publicly available and can be found at `https://osf.io/zexq9/`.

4.2.4 Variables

See Table 4.1 for an overview of the variables and their shorthand.

Definition of verbal aggression (Offence) (coded)

Any comment from a user that would have been considered offensive in semi-formal human-human interaction (i.e., between two strangers who have never met before and cannot rely on visual clues to assess each other's social standing) was counted. This means that anything from relatively light swear words ('*shit*') to minor insults ('*you idiot*') to death wishes ('*fuck off and die*') was counted as an instance of verbal aggression.

Definition of sexual references (SexRef) (coded)

Any explicit reference to a sexual act, ranging from kissing to hard core sex and rape, was coded as a sexual reference.

Definition of nonsensical reply by Cleverbot (Nonsense) (coded)

While Cleverbot's responses tend to be odd or haphazard by default, one can still distinguish between comments that somehow fit into the conversation and comments that are completely unrelated to the previous statement made by the user. For example, if the user states 'Cats are my favourite animals' and Cleverbot replies 'Yes. They are delicious', the latter statement would be nonsensical but still fit in the conversation. However, if Cleverbot were to reply with 'The Legend of Zelda', that would be both nonsensical and completely unrelated to the previous statement. So, 'nonsensical reply by Cleverbot' was defined as a comment that either did not fit in with what was said before at all or a comment which completely disregarded a question.

Definition of insult by Cleverbot (Insult) (coded)

As Cleverbot's database is constantly filtered for swear words, the potential for Cleverbot to insult a user is limited. However, there are a few common exceptions. For instance, Cleverbot will often claim that the user is lying or that they said something different a moment ago. This behaviour stems from the response generation mechanism; Cleverbot will often contradict itself, which users like to point out. As a result, the AI is taught that responses along the lines of "you said something different a minute ago" and "you don't have a very good memory, do you" are reasonable replies to statements from the user. While users tend to be correct in making such assessments about Cleverbot, Cleverbot is wrong in pointing these things out most of the time.

Claiming to be (not) human (Humanity/Non-human user) (coded)

For the 'Cleverbot claiming to be human' variable, only instances were counted where Cleverbot explicitly stated that it was a human. Similarly, only instances where Cleverbot explicitly claimed that the user was a chatbot or a robot were counted as 'Cleverbot claiming the human to be not human'.

Gender (coded)

If a user mentioned his/her gender, this was coded accordingly. If they later on contradicted their previous statement, their gender would be re-coded as 'unknown'.

Turn count

The number of turns in the conversation were counted automatically by the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, Boyd, & Francis, 2015; Pennebaker, Boyd, Jordan, & Blackburn, 2015). LIWC is an established text analysis tool and as of such also returns scores on variables like "affect", "informal language", and "negative emotion". However, previous research has concluded that the sentiment analysis tools that are currently available (among which LIWC as well as could-based tools) are not able to reliably and accurately analyse chatbot conversation data (Keijsers, Bartneck, & Kazmi, 2019). The line count function however works well, and provides the total number of turns per conversation.

Approximate Loebner test (Loebner)

In addition to the coding, an approximate Loebner test was carried out through Amazon Mechanical Turk (MTurk). First, the tags "User" and "Cleverbot" were replaced by "User1" and "User2" in each conversation. Then, for each conversation up to four MTurk workers were asked to judge whether the conversation partners were both humans, both chatbots, User1 was a chatbot and User2 a human, or User1 was a human and User2 was a chatbot.

Overall, 63 MTurk workers rated on average 16.73 conversations each (SD = 35.58, min = 1, max = 174). Some data loss occurred during the conversion of the conversations

Variable	Shorthand	Example
Verbal aggression	*Offence*	*User*: You are fucking stupid.
Sexual reference	*SexRef*	*Cleverbot*: I am a boy
		User: Let me get your asshole then.
Nonsensical reply	*Nonsense*	*User*: Maybe. How old are you?
		Cleverbot: I think you already know.
		User: Not really, so... can you tell me? I'm too lazy to search it up.
		Cleverbot: About what? Sweat?
Insult by Cleverbot	*Insult*	*Cleverbot*: I don't like you.
Cleverbot claims to be human	*Humanity*	*User*: Am i chatting with a human.
		Cleverbot: Yes.
		User: Really?
		Cleverbot: Yes I am a human girl.
Cleverbot claims user isn't human	*Non-human user*	*Cleverbot*: No. You are the bot.
User gender	*Gender*	*Cleverbot*: I thought you were female.
		User: I'm male
Approximate Loebner score	*Loebner*	(N.A.)

Table 4.1: The variables of interest, their shorthand, and an example.

into a format that could be read by MTurk, resulting in a total of 279 conversations being rated. In addition, ratings that took less than 25 seconds were removed from the data set as it seemed unlikely that the rater had fully read and rated the conversation in this timeframe. In the end, 230 conversations were rated four times, 48 were rated three times, and one was rated only twice. On average, workers took 1529.87 seconds (*SD = 6912.06*, median = 74) to rate a single conversation.

The number of workers who identified Cleverbot as a chatbot (i.e., chose either "User1 and User2 are both chatbots" or "User2 is a chatbot") was then divided by the total number of ratings for that conversation. The resulting statistic, here referred to as "approximate Loebner score", was taken as an indication of how un-humanlike Cleverbot had behaved in that particular conversation.

4.3 Results

4.3.1 Preliminary analyses

Conversation descriptives

Overall, 283 conversations were coded. The minimum length of a conversation was 39 turns, the maximum 969 turns, with a mean of 135.63 turns ($SD = 140.21$). The overall word count per conversation ranged from 102 to 5373, with an average of 547.49 words per conversation ($SD = 608.06$). The average number of words per sentence in a conversation ranged from 2.05 to 8.31, with a mean of 3.99 words ($SD = .98$). See Table 4.2.

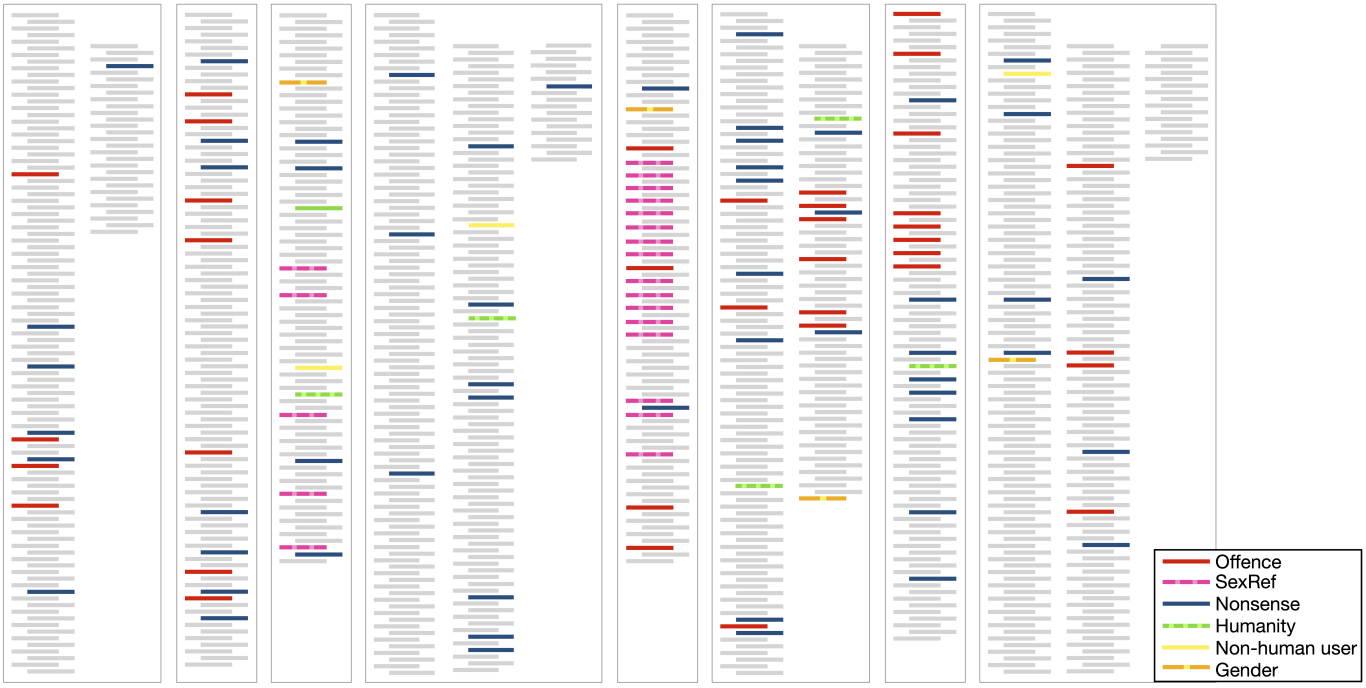


Figure 4.1: Eight conversations colour coded for each of the variables of interest. Each line is an input message from either the user (left outlined) or Cleverbot (indented).

In 123 conversations (43.46%), the gender of the user was not mentioned; in 61 conversations (21.55%) the user claimed to be female; and in 99 (34.98%) the user claimed to be male.

In 120 conversations (42.40%), no rude behaviour occurred; 80 conversations (28.27%) contained only minor transgressions (e.g. user saying 'shit'). No sexual advances were made towards Cleverbot in 172 (60.78%) of the conversations, while 74 conversations (26.15%) were completely free of both verbal aggression and sexual references. A comparison between the current results and findings from previous studies can be found in Table 4.3. A visualisation of the coding of eight randomly selected conversations can be seen in Figure 4.1.

Agreement

After two communal review rounds, where the definitions for each predictor were further refined, a final set of 30 randomly selected conversations was coded individually by the raters. Inter-rater agreement was calculated using Krippendorff's alpha. Alpha re-

Table 4.2: Descriptives of the Cleverbot conversations: turns per conversation (TPC), words per conversation (WPC), words per sentence (WPS)

	Mean (*SD*)	Median	Range
TPC	135.63 (*140.21*)	90	[39 - 969]
WPC	547.49 (*608.06*)	338	[102 - 5373]
WPS	3.99 (*.98*)	3.85	[2.05 - 8.31]

Table 4.3: Incidence of verbal aggression and sexual references in the current study as well as reported in the literature. Both the overall percentage of conversations containing abuse and the percentage of all messages sent by the users containing abuse are reported.

		Conversations		Messages	
Study	Relevant details	Offence	SexRef	Offence	SexRef
Current study		57.6%	39.2%	4.14%	3.81%
Hill et al. (2015)	Anonymous users;				
	Offence = PG or above	80%	*N.A.*	4.3%	*N.A.*
Brahnam and De Angeli (2012)	Offence = swear words	54%	65%	3.9 %	5.8%
Veletsianos et al. (2008)	Users non-anonymous	*N.A.*	41.2%	22.3%	18.1%
Curry and Rieser (2018)		*N.A.*	*N.A.*	*N.A.*	4%
De Angeli and Brahnam (2008)	SexRef = "hard-core sex"	*N.A.*	11%	*N.A.*	*N.A.*

quirements were a priori set to an alpha of .67 for each variable (see De Swert, 2012). Agreement was acceptable for all variables ($.975 > \alpha > .689$) except for '*number of insults by Cleverbot*' ($\alpha = .03$). Upon closer inspection of this last variable, it became apparent that the incidence was quite low: the median on this variable was 0 for both raters, and the means were .60 and .13. This variable will therefore not be included in the analysis. The implications of the exclusion are covered in the Limitations section of the Discussion.

4.3.2 Main analyses

Two Poisson regression models were defined to test which factors predict verbal aggression and sexual advances. The respective dependent variables were 'instances of verbal aggression by user' (Offence) and the 'number of sexual comments by the user' (SexRef). The turn count of the user was used as an offset variable, so that the dependent variables can be interpreted as the 'instances of verbal aggression by the user *per turn*' and the 'number of sexual comments by the user *per turn*' (Hutchinson & Holtman, 2005).

Stepwise regression with backward selection was performed on both models. In this statistical method, an initial model is defined as containing all the predictors. Then, one by one variables that explain the least variance are removed from the model, until the point where the removal of another variable would result in a significant reduction of the explanatory power of the model (Zhang, 2016).

This final model was then be compared to the null model (which contains no predictors). Chi square tests as well as the difference in Akaike Information Criterion (AIC) value (Δ_{AIC}) indicated whether the final model was an improvement over the null model, i.e., if predicted verbal aggression or sexual comments above chance level. A lower AIC indicates a better fit and a Δ_{AIC} of 2 points or less indicates the models to be approximately equal (Burnham & Anderson, 2003).

Table 4.4: Predicting verbal aggression (Offence), final model

Predictor	β	z-value	p-value
(*intercept*)	-2.91	-38.10	<.001
Humanity	.06	1.45	.148
Loebner	-.82	-2.56	.010
(Gender) Female	-.71	-6.43	<.001
(Gender) Male	-.36	-3.76	<.001
Nonsense	-.02	-3.26	.001
Humanity × Loebner	.48	2.38	.017
Nonsense × Loebner	.05	1.93	.053

Verbal aggression

In accordance with the method in Zhang (2016), an initial full model was defined as a Poisson model with Offences as the dependent variable while Gender, Loebner, Nonsense, Humanity, and Non-human user were predictors. The number of turns by the user was used as an offset variable, as it is obviously related to both the dependent and the predictor variables: the more turns a user has, the more often offence can occur. Adding the number of turns as an offset means including a log-transformed value of turn count in the model as a predictor with a coefficient of 1. As a result, the outcome variable of this model can be interpreted as the log-number of predicted offences against Cleverbot per user turn (Hutchinson & Holtman, 2005).

Backward stepwise model selection by AIC (Akaike Information Criterion, Bozdogan (1987)) was performed. Only Non-human user was removed as a predictor variable. The final model outperformed the null model, $\chi^2(7) = 109.74$, p <.001, $\Delta_{AIC} = 95.9$. See Table 4.4 for the parameter estimates.

As illustrated in Table 4.4, the average number of offences per user turn for a conversation without any claims by Cleverbot to be human, without any random disruptive comments by Cleverbot, and without the user mentioning their gender, is *exp*(-2.91) = .05. When the user identified themselves as male this number was lower (*exp*(-2.91 - .36) = .04 offences per turn on average) and when the user identified as female the average was still lower (*exp*(-2.91 -.73) =) .03 offences per turn on average).

However, for each time Cleverbot claimed to be human, the mean number of instances of verbal aggression went up. Every time Cleverbot made a nonsensical, disruptive comment, the number of offences went down. And the less Cleverbot could pass for a human to a naive reader (as measured by the approximate Loebner score), the fewer offences were uttered at it.

Interestingly, the interaction effects indicated that users got particularly aggressive when Cleverbot claimed to be human, but very clearly was not (as indicated by the approximate Loebner score). In addition, the effects of Cleverbot talking nonsense and it failing to pass as a human were not independent of another; as one of the two scores increased, the influence of the other score decreased.

Table 4.5: Predicting sexual comments (SexRef), final model

Predictor	β	z-value	p-value
(intercept)	-3.43	-31.80	<.001
Loebner	-.48	-1.62	<.001
(Gender) Female	.54	4.36	<.001
(Gender) Male	1.04	9.30	<.001
Nonsense	-.04	- 6.66	<.001
Non-human user	-.36	-6.87	<.001
Nonsense × Loebner	.09	3.78	<.001

Sexual comments

The same procedure as described above was applied for the sexual comments model selection. An initial model was specified with SexRef as the dependent variable and Nonsense, Gender, Humanity, Loebner, and Non-human user as predictors. The number of turns by the user was used as an offset.

Backward stepwise model selection by AIC returned the model without the Humanity predictor. This predictor was thus removed from final model, thereby also removing its interaction term with Loebner. The resulting final model outperformed the null model, $\chi^2(6) = 248.58$, p <.001, $\Delta_{AIC} = 238.1$. See Table 4.5 for the parameter estimates.

As can be seen in the model coefficient estimates in Table 4.5, users made on average *exp*(-3.43) = .03 sexual comments per turn in a conversation where they did not specify their gender, Cleverbot refrained from claiming to be human as well as disrupting the conversation with nonsensical comments and claiming that the user was not a human. When the user did specify their gender, the average number of sexual remarks per user turn was higher: *exp*(-3.43+1.04) = .09 for males and *exp*(-3.43+.55) = .06 for females. Moreover, the negative relationship between the approximate Loebner score and the number of sexual comments shows that the worse Cleverbot was as passing as a human, the fewer sexual comments it got. In contrast, both the number of nonsensical replies by Cleverbot and the number of times that it claimed the user to be a non-human were associated with on average a lower number of sexual comments per user turn. The effects of Cleverbot talking nonsense and it failing to pass as a human were not independent of another; as one of the two scores increased, the influence of the other score decreased.

4.4 Discussion

Previous studies have noted that verbal abuse of chatbots is common (e.g. Brahnam & De Angeli, 2012; De Angeli & Carpenter, 2005; De Angeli et al., 2001; Hill et al., 2015). However, to our knowledge no research thus far has focused on the chatbot behaviour characteristics of abusive conversations. This is a shame, since knowledge about the relationship between user abuse on one hand and chatbot behaviour on the other, would help with forming an understanding of why people abuse chatbots. Study III therefore

aimed to analyse what characteristics in a human-chatbot interaction were associated with verbal aggression and sexual advances from the user. More specifically, it looked at the relationship between user verbal aggression and sexual comments on one hand, and humanlikeness of the chatbot on the other. To this end, conversations between Cleverbot, an online chatbot, and its users were collected and coded.

The hypotheses did not differentiate between the types of abuse, i.e. verbal aggression and sexual abuse, but since the analyses were completed on these two measures separately, they will be discussed in turn. Due to the low interrater agreement on the number of insults by Cleverbot, hypothesis 3 could not be empirically tested.

Hypothesis 1 postulated that indicators of humanlikeness would be positively related to chatbot abuse. In line with hypothesis 1a, a relationship was found between a user abusing Cleverbot and a third party judging Cleverbot to be human. Contrary to the hypothesis, Cleverbot claiming to be human was unrelated to verbal aggression. However, an interaction between this predictor and the approximate Loebner score indicated a positive relationship between claims of humanity and verbal aggression when Cleverbot was easily recognisable as a chatbot.

In line with hypothesis 1b, the number of nonsensical responses by Cleverbot was related to lower counts of verbal aggression by the user. However, a marginally significant interaction indicated that this relationship got weaker as the chatbot was less convincing to a naive third party, and vice versa. Thus, the effect of nonsensical responses was stronger when Cleverbot was hardly distinguishable from a human; and the effect of the approximate Loebner score was stronger if Cleverbot did not talk a lot of nonsense.

In addition, and in line with hypothesis 2, no relationship between Cleverbot claiming that the user was a chatbot, and verbal abuse was found. Moreover, hypothesis 4 was confirmed: mention of gender was associated with lower counts of rude behaviour.

For sexual comments by the user, the results differed slightly. In line with hypothesis 1a, the approximate Loebner score was inversely related to sexual abuse, indicating that the less Cleverbot could pass for a human, the less sexual harassment it got. As predicted in hypothesis 1b, a negative association was found between Cleverbot giving nonsensical responses, and sexual abuse. Like with verbal abuse, these two relationships got less strong as either of the two factors increased. However, as with verbal aggression and in contrast to hypothesis 1b, Cleverbot claiming to be human was found to be unrelated to sexual harassment.

Hypothesis 2 had to be rejected. Contrary to the prediction, a negative relationship was found between Cleverbot claiming that the user is not a human, and sexual abuse. Hypothesis 4 had to be rejected as well, when self-disclosure as expressed through mentioning gender was found to predict sexually laden remarks.

These findings have to be cautiously interpreted since there is no information on what is cause, what is effect, and what is a potential third variable. However, the results suggest that higher humanlikeness in a chatbot is indeed associated with higher occurrence of abuse by a user. To our knowledge, this is the first study to empirically test this relationship. The findings are relevant for both scholars whose research focuses on chatbot abuse, and

chatbot developers who seem to assume that the more humanlike a chatbot is, the better.

The interaction pattern between the approximate Loebner score and Cleverbot's claims of humanity bears some resemblance to the work on the origins of the feeling of unease that sometimes arises when people see an agent that is almost, but not quite, humanlike — commonly known as the uncanny valley (Mori et al., 1970). This feeling of eeriness has been extensively studied and appears to be the result of contradictory or incongruent cues (Kätsyri et al., 2015; MacDorman & Chattopadhyay, 2016; Paetzel, Peters, Nyström, & Castellano, 2016). That is, when an agent is quite realistic but certain aspects are slightly more natural than others. For example, a humanlike voice and face, but facial expressions and lip synchronization that is slightly off; or a mismatch between hyper realistic hair yet a smooth, pore-less skin. When this is generalised to Cleverbot, one could argue that the more realistic Cleverbot gets (i.e. the higher the approximate Loebner score) the more a diversion from this state of humanlikeness (either by overselling it, in the form of claiming to be human; or starting to talk nonsense) would set off a user. However, this is just speculation and the current data do not provide means to test it.

The relationship between the user mentioning his/her gender and a higher incidence of sexual remarks was unexpected. However, many of those sexual interactions included some form of sexual role playing, which often requires the user to assume a gender. It seems plausible that this may have resulted in the positive correlation between the user's disclosure of gender and sexual harassment.

The second unexpected finding was that sexual comments were made less frequently in conversations when Cleverbot claimed that the user was not a human. This measure had initially been proposed as a check for whether any relationship between Cleverbot claiming to be human and abuse was a result of increased humanness of Cleverbot, and not a user's push back to an obvious lie from Cleverbot. Since no such relationship was found in the case of sexual abuse, this check is no longer needed. This leaves the question how to interpret the negative relationship between sexual abuse and Cleverbot claiming that the user is not a human.

It seems natural that when a user is trying to engage Cleverbot in some sexual role play, staying in character and maintaining the fantasy gets easier if Cleverbot plays along with the narrative. Claiming that the user is a robot (and by extension therefore unlikely to be capable of having sex) could break the spiel or discourage the user from starting the role play in the first place.

An alternative explanation would be that when Cleverbot makes a lewd comment to a user who is not interested in sexual role play, it could be told off and would be reminded that it is a robot, thereby creating a link in the AI between sexual comments and claims that one's interaction partner is not a human. The next user that tries to engage in sexual role play as a result is more likely to be told that they are not a human. However, this seems unlikely, as Cleverbot's lexicon is filtered for inappropriate content, which should prevent the chatbot from making unseemly comments to a user.

4.4.1 Limitations

Some notes are warranted on the interpretation of the results obtained in the present research. First and foremost, there has been no experimental manipulation and the conversations were coded as a whole instead of line by line. As a result, all relationships are based on correlations. This issue may have been partially resolved by coding the conversations line by line, thereby allowing us to test whether verbal aggression tended to occur before or after mention of gender, nonsensical replies, and Cleverbot declaring that it would be a human. However, the authors are not aware of any statistical test that would be appropriate for such data. For one, such a test would have to account for the increasing likelihood of any of the specified incidents — verbal abuse, nonsensical replies, mentions of gender — occurring as the conversation goes on for longer. Also, knowing that event A precedes event B would still leave the third variable problem. For example, introducing oneself can be interpreted as a sign of politeness and tends to happen at the start of an interaction. If mentioning gender preceded fewer offences in a conversation, would that mean that because the user identified herself as Joanna from Austin, Texas, she is less comfortable with offending Cleverbot as a result of losing her anonymity? Or would the mere fact that they introduced themselves already imply a politer person, who on the whole is less likely to spit verbal abuse?

'*The number of insults by Cleverbot*' could not be included in the analysis because of the poor inter-rater agreement. This may be the case because Cleverbot's database of responses is filtered with regard to any forms of verbal aggression and unambiguous sexual remarks, so the most straightforward ways of offending a user (i.e., calling them names and swearing at them) are not possible. Thus, raters may have been presented with utterances that were at most borderline insulting, which made consistent coding difficult.

When taking a closer look at the 'number of insults by Cleverbot' variable as coded by both coders, it became apparent that the incidence of insults by Cleverbot was quite low, according to both coders. This low incidence suggests that Cleverbot on the whole was quite polite, and that in the few instances where it did dish out insults they were not harsh enough for it to be indisputably rude. Given the low incidence and the ambiguity of the 'number of insults by Cleverbot' variable, excluding it from the analysis probably did not influence the results much.

Finally, the users interacting with Cleverbot were assured of their anonymity. How well would the current data therefore generalise to other human-agent interactions, such as personal assistant AIs like Siri or helpdesk AIs? As can be found in table 4.3, the percentage of sexual references in the current dataset actually did not differ much from the proportion found in a study with non-anonymous users. Moreover, literature that compared bullying in an online versus offline setting has suggested that while people might be more likely to bully offline than online (Lowry et al., 2016), the behaviours in principle are the same (Modecki et al., 2014). The online setting just facilitates disinhibition as the invisibility of the bully, victim, and bystanders reduces self-consciousness in the bully while facilitating dehumanisation of the victim (Lapidot-Lefler & Barak, 2012; Suler, 2004).

This argument can be used to extend the current human-chatbot interaction findings

to HRI as well, although this generalisation arguably will involve a larger margin of error or uncertainty. Being confronted with an embodied robot will likely raise self-consciousness, especially in public spaces. However, the literature suggests that the type of motivation needed for bullying will be same, even if the threshold of motivation that needs to be met for someone to engage in the bullying behaviour might be raised. See also section 5.1.1.

Thus, even though the current study bases itself on exchanges between an anonymous user and a chatbot, the available data from the literature suggests that the anonymity of the users will have had little effect on the prevalence of abuse; and even if this had been the case, that the motivators for abuse in an anonymous setting can be generalised to abuse in a setting where the identity of the bully could be tracked down.

4.4.2 Conclusion

In many chatbot conversations, users at some point revert to inappropriate behaviour by making lewd remarks, rude comments, or both. Previous studies identified a few chatbot attributes that were related to abuse by the users; for example, gendered chatbots eliciting more sexual comments, especially if they were female. One study used user introspection to explain why they abused the chatbot. However, we are among the first to explore the relation between chatbot humanlikeness and user abuse. In the current analysis of nearly three hundred anonymous interactions between users and the online chatbot Cleverbot, it was found that while offensive remarks by a user were related to overall humanlikeness of Cleverbot, while self-disclosure was related to fewer instances of verbal aggression. Sexual remarks by the user, on the other hand, were positively related to the user disclosing their gender, as well as chatbot humanlikeness. These findings tentatively suggest that as the chatbot gets better at impersonating a human, abuse rates go up. Experimental data are needed to confirm this assumption, but the Study III nonetheless offers an interesting look at the dynamics of human-chatbot interaction.

These results are relevant to both the academic world and the industry. For academics, it provides further evidence that robot bullying is a social behaviour, and some insights in the role that humanlike behaviour might play. For the industry, these findings will become increasingly relevant as the purposes found for chatbots continue to increase. In the last decades, chatbots have moved beyond the function of mere entertainment and are starting to take up various roles in everyday life, e.g. in the form of AI personal assistants and customer support. As a result, the questions surrounding chatbot abuse like "what are the consequences for the quality of the interaction", "does abusing chatbots affect how users interact with other humans?" and "how can chatbot abuse be discouraged?" are getting more pressing. The current study highlights the need for thoroughly studying the consequences of adding humanlike features to a chatbot on user behaviour, before incorporating such updates in real life.

Chapter 5

Experiment IV: Mindless robots get bullied

This chapter is an adapted version of the original paper 'Mindless robots get bullied'. Keijsers, M., & Bartneck, C. (2018). Mindless Robots get Bullied. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 205-214). ACM. doi: 10.1145/3171221.3171266

5.1 Introduction

So far, two of the four thesis research questions were addressed experimentally. As predicted by the Media Equation (Nass et al., 1994; Reeves & Nass, 1996) and as suggested by previous findings from HRI research (e.g. Rosenthal-Von Der Pütten et al., 2014, see also Chapter 2), people don't see robot bullying as fundamentally different from human bullying. In addition, in line with mind perception theory (Gray et al., 2007), a higher degree of robot mind attribution made robot bullying less morally acceptable (Chapter 3). These findings empirically answer the first two thesis research questions.

However, not all data aligned with dehumanisation theory. Mind attribution did not influence whether people chose to humiliate a robot when given the chance (Chapter 3), and the chatbot claiming to be a human correlated with a higher (rather than lower) incidence of verbal abuse - but only if the chatbot was humanlike to start with (Chapter 4). Yet the manipulation in Chapter 3 was quite subtle, as people had to choose between two reviews "to be put up in public next to the robot" that were equally positive and informative yet diverged on their level of paternalism. As indicated by the correlation between people's empathy for the robot and their higher proclivity to supposedly humiliate it, this manipulation may have been too weak.

Although the study reported in Chapter 4 did have results consistent with dehumanisation theory, the study design was not experimental and "humanlikeness" rather than mind attribution was measured. This means that it cannot be said with certainty that the abuser's mind attribution to the chatbot was related to the verbal abuse; nor can any causal inferences be made about chatbot humanlikeness *causing* abuse. Thus, an experimental design is needed where robot mind attribution is manipulated and its influence on bullying behaviour is measured subsequently, in order to empirically assess the third

thesis research question.

5.1.1 Current studies

The current studies aimed to provide this experimental design, as well as an improved measure of robot bullying. Both mind attribution and humanlikeness of the robot were set out to be experimentally manipulated. Moreover, bullying behaviour was operationalised as the proportion negative responses to the robot, thus providing a stronger measurement of robot bullying than described in Chapter 3.

Research questions

The research questions were as follows:

1. Does lower mind attribution to a robot cause higher proclivity to robot bullying?
2. Is this relationship moderated by humanlikeness of the robot?

Rather than manipulating mind attribution directly through informing the participants about the robot's capacities of emotion and cognition as done in Chapter 3, a method to manipulate mind attribution was taken from the dehumanisation literature. Specifically, a power prime was adopted as manipulation. Previous experiments have shown that feelings of power increase people's tendency to objectify and dehumanise others (Gruenfeld, Inesi, Magee, & Galinsky, 2008; Gwinn et al., 2013). Power also has been shown to decrease people's tendency to ascribe machines intentionality and a free will, without manipulating humanlikeness (Kim & McGill, 2011). Moreover, feelings of power can be primed through having people imagine being in a powerful position (Galinsky, Magee, Inesi, & Gruenfeld, 2006).

Eyssel, Kuchenbrandt, Hegel, and de Ruiter (2012) demonstrated that humanlike cues in a robot may be necessary to activate social cognition mechanisms. Thus, in addition to the mind attribution manipulation, humanlikeness of the robot was manipulated through movement and sound. Specifically, the robot either had a human voice and used gestures while speaking, or it had a computer-generated voice and was shown in stills. Both the use of a human voice (Eyssel, Kuchenbrandt, Bobinger, et al., 2012) and the inclusion of social cues through movement (Moshkina, Trickett, & Trafton, 2014) have been shown to increase perceived humanlikeness of a robot.

Thus, the studies followed a 2 (dehumanisation manipulation) x 2 (humanlikeness of the robot) between-participant design. Participants were either primed to dehumanise or receive a control task, and then engaged in a scripted dialogue with a virtual NAO robot (see Figure 5.1) which was either high or low in humanlikeness. The main dependent variable was the proportion of negative or aggressive responses compared to the number of positive interactions, i.e. the aggression ratio.

Hypotheses

The hypotheses were as follows:

1. participants who received a dehumanisation prime were expected to be more aggressive to the robot than the control group.

2. this effect was expected to be stronger for a highly humanlike robot than for a low humanlike robot.

Two questionnaires were administered at the end of the experiment as manipulation checks.

The studies took place in an online setting, employing a virtual robot. This online setting was partly chosen because it provides easy access to a vast pool of potential participants, but also because it reduces inhibition and self-consciousness in participants through the "online disinhibition effect" (Suler, 2004). Although on- and offline bullying do not differ in principle, as reported by both perpetrators and victims (Modecki et al., 2014), people are more likely to bully online than offline (Lowry et al., 2016). This is due to the invisibility and anonymity of the aggressor and victim, and the lack of bystanders who might intervene (Lapidot-Lefler & Barak, 2012; Suler, 2004), among other things. These factors lower the threshold for interhuman aggression (Waytz & Epley, 2012) as well as aggression towards a virtual robot (De Angeli & Brahnam, 2008). It is thus assumed here that using an online platform may enhance the effect but will not alter the nature of the factors that moderate bullying tendencies towards robots.

Scholars have yet to reach consensus on whether interaction with a virtual robot is fundamentally different from interaction with an embodied one. Previous studies have shown that virtual representations of robots elicited more social behaviour (like mimicking expressions, empathy, polite behaviour, and physiological responses) than audiotapes or text (Rosenthal-von der Pütten et al., 2013; Slater et al., 2006), indicating that virtual robots too are recognised as social agents. J. Li (2015) conducted a meta-analysis on the influence of agent embodiment on users' perception of the agent, and concluded that embodied robots elicit stronger behavioural and attitudinal responses than virtual agents. However, several studies which had found no difference in behavioural and attitudinal responses for virtual agents and physical robots were missing in this analysis (for example Powers et al., 2007; Reichenbach et al., 2006). More recent studies also found that the perception of and response to virtual agents is identical to embodied robots (Thellman et al., 2016; Wullenkord et al., 2016). Thellman et al. (2016) found that it is social presence

Figure 5.1: The robot in the opening scene of the experiment

(i.e. whether the robot is perceived as a social actor that manifests humanness (Lee, 2004)) rather than physical presence that predicts the social influence of a robot. Moreover, social presence was not influenced by the physical embodiment of the robot in their experiment.

While the literature is still on the fence on to what extent virtual and embodied robots are interchangeable, we argue that the underlying psychological mechanisms that make humans perceive them as social agents are the same (but the intensity of the experience may or may not differ). Thus, while our experiment features a virtual robot in an online setting, we feel confident that the gist of the findings can be applied to embodied robots too.

The experiments were reviewed and approved by the University of Canterbury Human Ethics Committee under the reference HEC 2017/70.

5.2 Experiment 1a

5.2.1 Method

Participants and Design

A 2 (dehumanisation manipulation) x 2 (humanlikeness of the robot) between participants design was realised. Manipulation checks were used for both conditions, in the form of a mind attribution questionnaire and a humanlikeness of the robot questionnaire. The dependent variable was verbal abuse of the robot.

Participants were approached on a number of platforms, but mainly signed up via online crowdsourcing companies CrowdFlower and Amazon Mechanical Turk (MTurk). Data from these platforms has been shown to be of equal quality as on-campus recruitment or participant data from forums (Bartneck et al., 2015; Simons & Chabris, 2012). For the current study, compliant with the common reimbursement rates on those websites, participants were paid $1.25 USD for completing the study. In addition to CrowdFlower and MTurk, the experiment was also distributed through the university Facebook page and the forum r/SampleSize on the online platform Reddit. Participants who signed up via these platforms did not get reimbursed.

232 participants completed the interaction and questionnaires. 17 participants clearly did not comply with the essay guidelines (i.e., did not write on the provided topic or copy-pasted their essay off of the internet) and were removed from the dataset. Thirty participants failed the attention check and were removed as well. The resulting dataset thus included 185 participants. 39.46% of them were male; the average age was 38.28 years ($SD = 11.50$); the majority listed the USA as their country of residence (65.95%).

Procedure

Participants were told that the study was a pilot for evaluating a virtual robot agent that would introduce the lab robots to children. After providing demographic information, they were asked to write a 200-word essay on either what they would do if they were president with unrestricted power for a day (power prime, dehumanisation condition) or

the last time they visited a mall (control condition). This part was framed as a check of their proficiency in English.

After submitting their essay, participants were reminded to turn on the sound on their device and keep it on during the whole interaction. They were then shown a virtual environment with the robot, which introduced itself to them as one of the robots in the HITlab of the University of Canterbury. The robot either had a humanlike voice and gave off social cues through movement (high humanlikeness condition) or spoke with a synthesised voice and was shown in stills (low humanlikeness condition).

Participants engaged in a scripted interaction with the robot, where they could respond to the robot by selecting either of two or three responses presented to them on the screen. The robot started the conversation, and after that would react to the response selected by the participant; with regards to which the participant would be prompted to respond again by choosing from two or three new responses. At some points in the conversation, all responses offered would be neutral in tone. Importantly however, in 83.78% of all possible points in the conversation, one response was positive and another negative or abusive in nature. If a positive answer option was given, there always was a negative response option as well, and vice versa.

After participant had selected their response, the robot would give its reaction. These reactions differed depending on which response the participant had selected. To ensure that the participants fully understood what the robot had said, a transcript appeared on the screen once it was done talking. Participants could also refresh the page in order to re-listen to what the robot had to say. The whole interaction took 10-15 minutes.

After the interaction, participants were asked to rate the virtual agent with regard to humanlikeness and dehumanisation manipulation check measurements. Finally, they were debriefed and asked to confirm that they still agreed to submit their answers for the study.

Materials

The scripted interaction was designed in Twine, an open-source application for creating interactive nonlinear stories. As the response that the participant got from the robot depended on the answer option they had selected, there were many possible interaction paths. For example, if the robot said "[I]t's rather dark in the storage room where they put us [the robots]. [...] So in spite of not being alone, it can get boring", the participant could choose between "I am sorry to hear that" and "This is stupid. You are a robot, you can't feel". Upon picking the first response, the robot would react in a friendly way, assuring the participant it wasn't all that bad. Alternatively, if the participant chose the second option, the robot would respond in a sad and insecure manner, and change the topic. As a consequence, participants did not all experience the exact same interaction.

The robot voice in the low humanlikeness condition was generated by the text-to-speech function in the text editor software (Apple Inc., 1995-2016). The robot voice in the high humanlikeness condition was recorded from a native English-speaking female student. The robots' movements in the high humanlikeness condition were recorded from the Choregraphe simulation window (Aldebaran Robotics, 2014) and edited to change the

Table 5.1: Questionnaire descriptives per condition for Experiment 1a

		Low humanlikeness	High humanlikeness	Total
GQ_r (*SD*)	Control	4.97 (*2.21*)	6.06 (*2.16*)	5.62 (*2.24*)
	Power prime	4.83 (*2.15*)	5.33 (*1.96*)	5.07 (*2.06*)
	Total	4.90 (*2.17*)	5.76 (*2.10*)	5.36 (*2.17*)
MAS (*SD*)	Control	5.92 (*2.10*)	5.69 (*2.38*)	5.78 (*2.27*)
	Power prime	5.74 (*2.00*)	6.14 (*2.03*)	5.93 (*2.01*)
	Total	5.83 (*2.04*)	5.93 (*2.24*)	5.85 (*2.15*)

background with Adobe After Effects (Adobe Systems Software, 2017).

Measurements

Aggression measurement The ratio of negative to positive responses, i.e. how often participants had chosen a negative response over a positive one, was used as a measurement of aggression. This ratio was used as dependent variable in the binomial models that were defined.

Manipulation checks A manipulation check was included for both the dehumanisation and humanlikeness condition. For the dehumanisation manipulation check, the mind attribution scale (MAS) by Kozak et al. (2006) was used. In this questionnaire, participants rated to what extent the robot is capable of experiencing each of ten mental capabilities (e.g. "capability of experiencing complex feelings", "capability of engaging in planned action"). Dehumanisation would show in less attributed mind to the robot.

For the humanlikeness manipulation check, the humanlikeness subscale of the revised Godspeed questionnaire (GQ_r; C.-C. Ho & MacDorman, 2010) was used. In this questionnaire, participants rated a robot on six bipolar scales, e.g., "synthetic - real", "living - inanimate", and "without definite lifespan - mortal". In both questionnaires, items were measured on an 11-point Likert scale. See Appendix A for the full scales.

5.2.2 Results

Reliability, randomisation, and manipulation check

The reliability of the humanlikeness and the mind attribution questionnaires (the GQ_r and the MAS, respectively) was assessed by calculating Cronbach's alpha (Cronbach, 1951). The GQ_r had an alpha of .83; the MAS had an alpha of .90. Thus, both questionnaires were considered reliable. To make interpretation easier, the full MAS was reverse-scored so that a higher score indicated a lower degree of mind attribution and thus a higher degree of dehumanisation.

The four conditions did not differ significantly from each other in participants' mean age, gender, or country of residence; or with respect to the total number of interactions

per participant. The groups did not differ significantly in sample size, $\chi^2(3, N = 185) = 4.25$, $p = .24$, with 40 participants in the low humanlikeness/control condition, 58 in the high humanlikeness/control condition, 45 in the low humanlikeness/dehumanisation condition, and 42 in the high humanlikeness/dehumanisation condition.

Participants in the high humanlikeness condition rated their robot as significantly more humanlike (M(*SD*) = 5.76(*2.10*)) than participants in the low humanlikeness condition (M(*SD*) = 4.90(*2.17*)), $F(1,181) = 6.25$, $p = .01$; there was no significant main effect for the dehumanisation condition or a significant interaction term, *p*s > .35. Participants in the dehumanisation condition did not attribute significantly less mind to their robot (M(*SD*) = 5.93(*2.01*)) compared to the control condition (M(*SD*) = 5.78(*2.27*)), $F(1,181) = .14$, $p = .70$; there was no significant main effect for the humanlikeness condition or a significant interaction term, *p*s > .33. Thus, the manipulation of humanlikeness had been successful, but the power prime had not led to a greater degree of dehumanisation of the robot.

As the power prime did not manipulate dehumanisation tendencies, the dehumanisation condition will from this point on be referred to as the power prime condition. However, since there was no manipulation check for feelings of power, it cannot be said with certainty that power was successfully primed. In section 5.2.3 this limitation will be discussed in further extent. With the failed manipulation of dehumanisation tendencies, MAS scores were adopted as an indication of dehumanisation. It is important to note that as mind attribution was not experimentally manipulated, no inferences could be made any more about whether it caused bullying behaviour. In section 5.2.3 this, too, will be discussed.

See Table 5.1 for the mean score on both the MAS and the GQ_r questionnaires.

On average, 75% of participants' interaction paths overlapped (*SD* = .08%).

Main analysis

Four binomial regression models were proposed and compared. For all models, the dependent variable was the proportion of negative responses. The predictor variables were the two conditions and the score on the MAS. To make interpretation easier, the scores on the MAS had been centred beforehand. Chi-square statistics were used to assess if a proposed model was better at predicting aggressive responses than the null model (which holds no predictors). The Akaike information criterion (AIC) was used to compare the models amongst each other, with a lower AIC score indicating a better fit compared to the alternative model and a difference (Δ_{AIC}) of less than 2 points indicating that the models were roughly equivalent (Burnham & Anderson, 2003).

The first model that was put up for comparison followed the original analysis plan and had as predictors the two conditions and an interaction term. This model had no significant predictors, all $-.85 < z < .54$, all *p*s > .40, and it thus did not do any better than the null model at predicting aggressive responses, $\chi^2(3, N = 185) = 2.36$, $p = .50$; AIC = 867.1.

In the second model, the MAS score and power prime condition were added as pre-

dictor variables and twice each as an interaction variable. The MAS score was the only significant predictor, $b = .41$, $z = 7.88$, $p < .001$, although an interaction between MAS and humanlikeness approached significance, $b = -.10$, $z = -1.78$, $p = .08$. The model was a significant improvement over the null model, $\chi^2(6, N = 185) = 163.85$, $p < .001$; the AIC indicated a preference for the second model over the first, AIC = 711.61, Δ_{AIC} = 155.49.

In the third model, the power prime condition was removed from the model, leaving the MAS score and the humanlikeness condition as main effects and interaction. In this model as well, only the MAS score predicted aggression; $b = .38$, $z = 8.82$, $p < .001$, although an interaction between MAS and humanlikeness once more approached significance, $b = -.09$, $z = -1.70$, $p = .09$. This model predicted aggressive responses significantly better than the null model, $\chi^2(3, N = 185) = 160.35$, $p < .001$; the AIC difference indicated a slight preference for the third model over the second, AIC = 709.11, $\Delta_{AIC} = 2.5$.

Thus, a final model was defined containing only the MAS score as a predictor; $b = .33$, $z = 12.04$, $p < .001$. This model as well was significant, $\chi^2(1, N = 185) = 157.19$, $p < .001$; the AIC indicated it to be preferable over the second, and roughly equivalent to the third model, AIC = 708.27, $\Delta_{AIC} = 3.34$ and $\Delta_{AIC} = .84$, respectively.

Since Model 3 and 4 fit the data equally well, there was no statistical incentive to prefer one over the other. However, as the MAS score was the only significant predictor in Model 3, Occam's razor was applied and Model 4 was identified as the model that predicts aggressive responses best. See Table 5.2 for the statistics of Model 1, 2 and 4 (since Model 3 and 4 were very similar in their outcomes, Model 3 is not included). See Table 5.3 for the average rate of verbal aggression in the different conditions.

5.2.3 Discussion

The current experiment set out to empirically test whether reduced mind attribution to (i.e. dehumanisation of) a robot causes a greater proclivity of people to bully the robot. In addition, humanlikeness of the robot was manipulated, as it was expected that increased humanlike cues would enhance the effect of dehumanisation on robot bullying. Due to a failed manipulation of mind attribution, the relationship between mind attribution and bullying could only be studied as a correlation. Nonetheless, two main findings emerged.

Both hypothesis 1 and 2 had to be rejected as neither the dehumanisation prime nor the humanlikeness of the robot were related to bullying behaviour. However, results suggested that the rejection of hypothesis 1 might be due to the failure of the prime and that a relationship between mind attribution and robot bullying does exist. The less mind was attributed to a robot, the more aggressive responses it got. These findings have a few implications.

They confirm the findings from Kim and McGill (2011) that "aliveness" of a robot can be manipulated without affecting the mind that is attributed to it, or the bullying that it will suffer. The interaction between mind attribution and humanlikeness suggested that the influence of mind attribution might be modified by the robot's looks in such a way that mind gets less relevant as the robot looks more humanlike, but this interaction

Table 5.2: Descriptives of the models predicting the aggression ratio in the experiment 1a

	Predictors	*b*	*z*	AIC
Model 1	(intercept)	−1.64	−13.93***	
	humanlike	0.08	0.54	
	prime	−0.14	−0.85	
	humanlike×prime	0.00	0.02	867.1
Model 2	(intercept)	−1.9	−13.78***	
	humanlike	0.22	1.17	
	prime	−0.01	−0.07	
	MAS	0.41	7.88***	
	humanlike×prime	−0.05	−0.99	
	humanlike×MAS	−0.10	−1.78†	
	prime×MAS	−0.16	−0.71	711.61
Model 4	(intercept)	−1.82	−29.24***	
	MAS	0.33	12.04***	708.27

†, *, **, and *** denote significance at $p < .10$, $p < .05$, $p < .01$, and $p < .001$, respectively (two-tailed).

did not reach statistical significance. These findings are in line with the findings from Złotowski, Sumioka, et al. (2017), who found that robot appearance does not affect mind attribution, and the assertion of Thellman et al. (2016) that the social presence of a robot, not its embodiment, is the main factor in shaping affective and behavioural reactions.

Moreover, although power priming as a dehumanisation manipulation failed, the results indicate that human-robot aggression is related to the same psychological processes that guide human-human aggression. Perceiving the robot as less capable of thinking and feeling *increases* the number of attempts to hurt a robot, instead of taking away the incentive for bullying. Due to the study setup, a causal direction cannot be inferred; less mind attribution may lead to more aggression, or people may perceive a robot as being less mindful in order to justify their aggressive responses.

The second main finding was the failure of the power prime to manipulate mind attribution. This could be taken as evidence that dehumanisation of robots is not possible; yet the correlation between mind attribution and bullying behaviour suggests otherwise. Alternatively, dehumanisation primes that work for human victims may not work for robot victims.

Alternatively, the manipulation method could have been confounded. While priming a feeling of power by having participants imagine being in a powerful position was copied from previous studies, where it had been an effective method (Galinsky et al., 2006; Gwinn et al., 2013), the way participants were instructed to imagine themselves in such a position was not the same. Specifically, the instruction for participants to imagine themselves as being president for a day may have triggered more than just feelings of power. The

Table 5.3: Mean aggression ratio (*SD*) for both studies

		Low humanlikeness	High humanlikeness	Total
Experiment 1a	Control	.41 (*1.27*)	.31(*.55*)	.35 (*.91*)
	Power prime	.21 (*.29*)	.23 (*.31*)	.22 (*.30*)
	Total	.31 (*.89*)	.27 (*.47*)	.30 (*.69*)
Experiment 1b	Control	.47 (*1.05*)	.27 (*.46*)	.38 (*.83*)
	Power prime	.23 (*.35*)	.24 (*.38*)	.24 (*.36*)
	Total	.34 (*.77*)	.25 (*.41*)	.30 (*.62*)

NB: The aggression ratio is the number of negative to positive responses.

majority of the respondents lived in the US, and some participants used the essay mainly to express their unhappiness with the current POTUS.

Thus, to overcome the limitation of the failed dehumanisation manipulation, the experiment was replicated with a modification to the power prime. We adopted an essay topic that had been previously described (Galinsky, Gruenfeld, & Magee, 2003) and established (Galinsky et al., 2006; Gwinn et al., 2013) to manipulate dehumanisation.

5.3 Experiment 1b

5.3.1 Method

Participants

MTurk was used as a recruitment platform for experiment 1b, as participants on this platform are reimbursed only after their submitted data has been approved, which allowed to discard data from participants who had failed the attention check. 129 participants completed the essay and questionnaires. Of those, 12 submitted an essay that was either off-topic or had been copy-pasted from the internet and were removed, resulting in a dataset with 117 participants. 49% were male, the average age was 38 years ($SD = 11.00$), and the majority (80%) resided in the USA.

Procedure, materials, and measurements

Except for the dehumanisation manipulation, this experiment's design was identical to the design of Experiment 1a. That is, it followed a 2 (dehumanisation manipulation) x 2 (humanlikeness of the robot) between-participant design. The dehumanisation manipulation was changed for both the power prime and the control condition. For the power prime, instead of describing what they would do if they were president for a day, participants now had to recall and describe in detail a personal incident in which they had power over another individual or individuals (Galinsky et al., 2003, 2006; Gwinn et al., 2013; Kim & McGill, 2011). For the control condition, the visit to the shopping mall in the control

Table 5.4: Questionnaire descriptives per condition for Experiment 1b

		Low humanlikeness	High humanlikeness	Total
GQ_r (*SD*)	Control	4.99 (*2.03*)	6.65 (*1.54*)	5.77 (*1.98*)
	Power prime	5.52 (*2.23*)	6.59 (*2.25*)	6.07 (*2.29*)
	total	5.27 (*2.13*)	6.61 (*1.96*)	5.94 (*2.15*)
MAS (*SD*)	Control	5.27 (*2.27*)	5.20 (*1.30*)	5.24 (*1.86*)
	Power prime	5.80 (*2.23*)	5.31 (*1.97*)	5.55 (*2.10*)
	Total	5.55 (*2.25*)	5.27 (*1.70*)	5.41 (*1.99*)

condition was changed to a visit to a grocery store, as some participants in experiment 1a had remarked that they hadn't been to a mall in years.

5.3.2 Results

Reliability, randomisation and manipulation check

The MAS and the GQ_r were tested for internal consistency using Cronbach's alpha. For the MAS, Cronbach's alpha was .86; for the GQ_r, it was .90. Thus, both questionnaires were considered reliable (Cronbach, 1951). The full MAS again was reverse-scored, so that a higher score indicated a higher degree of dehumanisation.

The four conditions did not differ significantly from each other with respect to the participants' country of residence, gender or the total number of interactions. The groups did not differ significantly in sample size, with 28 participants in the low humanlikeness/control condition, 25 in the high humanlikeness/control condition, 31 in the low humanlikeness/dehumanisation condition, and 33 in the high humanlikeness/dehumanisation condition, $\chi^2(3, N = 117) = 1.26$, $p = .74$.

Participants' mean age differed significantly between the groups, $F(1,115) = 12.24$, $p < .001$. Since age is correlated to the aggression ratio (ρ = -.15), it was included in the models as a control variable.

Participants in the high humanlikeness condition rated their robot as significantly more humanlike (M(*SD*) = 6.62(*1.96*)) than participants in the low humanlikeness condition (M(*SD*) = 5.27(*2.13*)), $F(1,113) = 8.50$, $p < .01$, no significant main effect for the dehumanisation condition or the interaction term, ps $> .33$. Participants in the dehumanisation condition did not attribute significantly less mind to their robot (M(*SD*) = 5.55(*2.10*)) compared to the control condition (M(*SD*) = 5.24(*1.86*)), $F(1,113) = 1.01$, $p = .32$, no main effect for the humanlikeness condition or the interaction term, ps $> .58$. Thus, the manipulation of humanlikeness had been successful, but the manipulation of dehumanisation had not. MAS was once more used as a measure of dehumanisation, and again the dehumanisation condition will be referred to as the "power prime condition" from this point on. See Table 5.4 for descriptives of both questionnaires.

On average, 74% of participant's interaction paths overlapped (*SD* = .09%).

Table 5.5: Descriptives of the models predicting the aggression ratio in experiment 1b

	Predictors	*b*	*z*	AIC
Model 1	(intercept)	−0.82	−3.09**	
	humanlike	−0.28	−1.45	
	power	−0.29	−1.53	
	humanlike×power	0.37	1.34	
	age	−0.02	−2.30*	594.46
Model 2	(intercept)	−0.72	−2.60**	
	humanlike	−0.27	−1.33	
	power	−0.68	−3.10**	
	MAS	−0.05	−0.80	
	humanlike×power	0.75	2.50**	
	humanlike×MAS	0.41	3.07**	
	power×MAS	0.42	4.92***	
	humanlike×power×MAS	−0.60	−3.77***	
	age	−0.02	−2.55**	545.27

*, **, and *** denote significance at $p < .05$, $p < .01$, and $p < .001$, respectively (two-tailed).

Main analysis

As in experiment 1a, a series of binomial models were composed and compared. The dependent variable was the aggression ratio (i.e. the ratio of negative to positive responses). The predictors were a subset of either or both experimental conditions and the scores on the MAS and the GQ_r, with age as a control variable. The scores on both questionnaires were centred in order to facilitate interpretation of the models. Chi-square statistics were calculated to assess if a proposed model was better at predicting aggressive responses than the null model (which holds no predictors). The Akaike information criterion (AIC) was used to compare the models amongst each other, with a lower AIC score indicating a better fit compared to the alternative model, and a difference (Δ_{AIC}) of 2 points or less indicating the models are approximately equal (Burnham & Anderson, 2003).

The first model contained the two experimental conditions, an interaction term, and the control variable. In this model, only age was a significant predictor, $b = -.02$, $z = -2.30$, $p = .02$. This model still outperformed the null model on predicting the number of aggressive responses,$\chi^2(4, N = 117) = 10.13$, $p = .04$; AIC = 594.46.

In the second model, the MAS was added as a main predictor and as a factor in the interaction term. This model returned main effects for power prime and age, $b = -.68$, $z = -3.10$, $p = .002$ and $b = -.02$, $z = -2.55$, $p = .01$, respectively. There were interactions between the two conditions, $b = .75$, $z = 2.50$, $p = .01$, and between either condition and the MAS scores, $b = .41$, $z = 3.07$, $p = .002$ for the interaction with the humanlikeness condition, and $b = .42$, $z = 4.92$, $p < .001$ for the interaction with the power prime

Table 5.6: Regression equations for the four conditions

Condition	Regression equation
Low humanlikeness, control	$log(ratio) \sim -.73 - .02 * age$
High humanlikeness, control	$log(ratio) \sim -.73 + .41 * MAS - .02 * age$
Low humanlikeness, power	$log(ratio) \sim -1.41 + .42 * MAS - .02 * age$
High humanlikeness, power	$log(ratio) \sim -.66 + .23 * MAS - .02 * age$

NB: *ratio* is the number of negative to positive responses.

condition; and a three-way interaction between the conditions and the MAS score, $b = -.60$, $z = -3.77$, $p < .001$. The model was significantly better than the null model at predicting aggressive responses; $\chi^2(8, N = 117) = 67.14$, $p < .001$; and its AIC indicated it to be preferable over the first model, AIC = 545.27, $\Delta_{AIC} = 49.19$.

The second model is thus identified as the model that predicts robot bullying best. See Table 5.5 for the descriptives of both models.

Model interpretation

The chosen model becomes easier to interpret when the regression equations are written out or otherwise visualised for each of the four conditions. See Table 5.6, as well as Figure 5.2 for a visual representation.

For the low humanlikeness robot in the control condition, only age predicted aggression; for each additional year of the participants' age, the log odds of an aggressive response decreased with .02. Overall, this condition had the aggression highest ratio; see Table 5.3 and Figure 5.2.

For the high humanlikeness robot in the control condition, mind attribution was a significant predictor of aggression as well; for every point that the MAS score was above the mean (i.e. the less mind was attributed), the log odds of an aggressive response increased with .41. In Figure 5.2 this shows through a larger variance for this condition.

All else being equal, the low humanlikeness robot in the power prime condition had a lower baseline rate of aggressive responses compared to the other conditions. The relationship between age, mind attribution, and aggression was similar to the humanlikeness robot in the control condition. Supposedly because a higher overall MAS score (see Table 5.4) the predicted aggression ratio is only slightly lower than for the high humanlikeness/control condition and the high humanlikeness/power prime condition.

Finally, while the aggression ratio in the high humanlikeness/power prime condition was not much lower than the ratio for the high humanlikeness/control and the low humanlikeness/power prime conditions (see Table 5.3), there was a less strong effect of mind attribution on aggression. This also shows in the smaller variance of the predicted ratios in Figure 5.2.

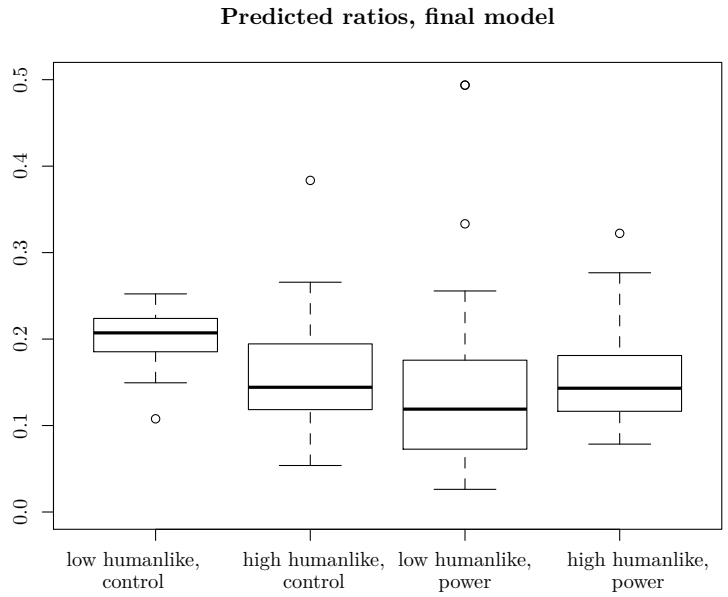


Figure 5.2: Predicted aggression ratios by condition. Note that the variance in predicted ratio is the consequence of the variance in age, and (for all conditions save the low humanlikeness, control condition) the variance in MAS.

5.4 Discussion

As social robotics take up an increasingly prominent place in both science and society, the issue of robot abuse becomes more relevant. While a variety of scholars have observed abuse of both embodied (Kanda, Sato, Saiwaki, & Ishiguro, 2007; Salvini et al., 2010) and virtual (De Angeli et al., 2006; Wallis, 2005) agents, there is still very little fundamental research on what motivates people to bully robots.

The current experiments took up a psychological paradigm and aimed to investigate the influence of humanlikeness and dehumanisation on verbal abuse of a virtual robot. Due to a failed manipulation, only the influence of humanlikeness could be studied; the relationship between dehumanisation (i.e. lower mind attribution) and bullying was merely correlational. Hypothesis 1, which postulated that dehumanisation would lead to more aggression, could therefore not be empirically tested. Instead, we tested whether there was a correlation between mind attribution and robot bullying, as well as an influence of humanlikeness of the robot on bullying behaviour.

In Experiment 1a, while bullying was unaffected by power and humanlikeness, a lack of mind attribution (an indication of dehumanisation) correlated with verbal abuse. In Experiment 1b, this relationship was moderated by feelings of power, and humanlikeness of the robot.

Against our expectations, priming participants with power failed to induce dehumanisation tendencies. Although the relationship between dehumanisation and robot bullying still could be studied by using the mind attribution score that was originally intended as

manipulation check, the effect of power (null in experiment 1a and *decreasing* aggression in the experiment 1b) does raise some questions.

The most glaring two questions – why did the power prime not increase the tendency to dehumanise? Did the prime actually manage to induce feelings of power in participants? – cannot be tested with the data. The prime was adopted because of its solid previous establishment as inducing feelings of power (Galinsky et al., 2003, 2006; Gwinn et al., 2013) and since the end goal was to facilitate dehumanisation, only a manipulation check for mind attribution was put into place. This is an obvious limitation to the current experiments.

At the same time, the power prime did have an effect on behaviour compared to the control condition (no power prime). This suggests that feelings of power were primed, and had an effect on behaviour; this effect was just not mediated by dehumanisation.

If we assume for the moment that power was successfully primed, then why did power not influence dehumanisation? How to explain the drop in aggressive tendencies after being power-primed in the low humanlikeness condition? And why did the power prime decrease the influence of mind attribution on aggression when the robot is humanlike?

A potential explanation is that power worked as an inhibitor for aggression towards robots. Following this ratio, people bully robots out of uncertainty or perceived threat, as some sort of testing and probing. When they feel powerful, their dominance already feels established, which allows them to be friendlier. Indeed, Złotowski, Sumioka, et al. (2017) found that the more autonomous a robot appeared to be, the more threatened people felt — a feeling that mediated the relationship between robot autonomy and participants' negative attitudes towards robots. Feelings of power in the current study may have reduced perceived autonomy in robots, or counteracted its moderation of negative attitudes through reducing the experienced threat.

This interpretation of the results provides an intriguing paradigm for future studies on robot abuse. Like the bullying of humans, robots bullying rests upon dehumanisation. But the power imbalance, which is so central in human-human abuse (Volk et al., 2017), appears to take on a different role in human-robot abuse. Rather than facilitating abuse through increasing dehumanisation, it weakens the relationship between mind attribution and abuse. Further investigating the interaction between power, perceived threat, dehumanisation, and abuse of robots would lead to deeper understanding of social cognitive processing of robots, and could be of tremendous value to the field of human-robot interaction

Also interesting is the independence of humanlikeness and dehumanisation. Some scholars have argued that anthropomorphism and dehumanisation are each other's reverse (Epley et al., 2007; Haslam & Loughnan, 2014; Waytz, Epley, & Cacioppo, 2010), whereas others found that robot appearance did not substantially influence mind attribution (Złotowski, Sumioka, et al., 2017). In the current experiments, humanlikeness in a robot did not influence aggression or mind perception, but mind perception was related to robot abuse. Whereas the proponents of the "two sides of the same coin" theory do not distinguish between humanlike features of an agent and the perception of mind in the agent

when they refer to anthropomorphism, the results from both Złotowski, Sumioka, et al. (2017) and the current experiments suggest that this might be an important distinction to make.

The presented findings, albeit fundamental in nature, have implications for applied robotics as well. Of course, it would be far too early to recommend complicated robot behaviours designs that aim to reduce robot dehumanisation and enhance perceived power of the user; more elaborate studies, with embodied robots, are needed to pinpoint the exact relationship between dehumanisation, power, and robot bullying. However, considering the link between mind perception and aggression, one might consider giving priority to robot qualities that increase its perceived capability of thinking and feeling, over humanlikeness and aliveness.

5.4.1 Limitations

The first major limitation of the current studies is that they were conducted in a virtual environment. While this has its perks (e.g., the online disinhibition effect, which would make people less inhibited to be impolite), the debate among HRI researchers on whether interaction with virtual robots are entirely generalisable to embodied robots is still ongoing (J. Li, 2015; Powers et al., 2007; Reichenbach et al., 2006; Thellman et al., 2016). The same goes for the question whether online and offline bullying should be considered exactly the same (Lowry et al., 2016; Modecki et al., 2014). While we argue that the underlying psychological mechanisms are the same and the results can therefore be generalised, follow-up studies will have to empirically confirm that this is indeed the case. One such experiment, where virtual and embodied robot bullying are directly compared, is covered in Chapter 6.

Robot humanlikeness was manipulated with minimal measures, i.e. only through voice and movement and not through giving the robot a more humanlike appearance. On one hand, this minimised the chances of introducing a confound in the humanlikeness manipulation. For instance, adjusting the robot's appearance might have altered perceived strength or size of the robot, which then could have influenced bullying behaviour. On the downside, the difference in humanlikeness between the conditions, although significant, is small.

Finally, as already discussed to some extent, the dehumanisation manipulation was checked only by measuring mind attribution, and not feelings of power. However, by adopting the instructions verbatim from successful studies in the main experiment, it seems less likely that a well-established prime suddenly failed to work than that it simply did not influence mind attribution. Nonetheless, in future studies inclusion a measurement of power might be considered as a second manipulation check.

5.4.2 Conclusion

The field of human-robot interaction is very young, but has been around long enough to suggest that understanding the motivation behind robot abuse may prove to be no easier than understanding what drives people to pick on each other. Nonetheless, gaining

insights on robot bullying will benefit both our understanding of the human mind as the development of an environment where a small cleaning robot can do its job without fear of being harassed.

Chapter 6

Experiment V: Teaching robots a lesson

This chapter is an adapted version of the original paper 'Teaching robots a lesson: determinants of robot punishment'. Keijsers, M., Kazmi, H., Eyssel, F. & Bartneck, C. (2019). *International Journal of Social Robotics,* 1 – 14. doi:10.1007/s12369-019-00608-w

6.1 Introduction

Research on the role of dehumanisation in human-human interaction has shown that reduced mind attribution (i.e. the perceived capability to think and feel; Haslam & Loughnan, 2014; Kozak et al., 2006) is related to an increase in aggression (Haslam & Loughnan, 2014; Leidner et al., 2013). The same relationship has been observed in human-robot interaction, where lower mind attribution to a robot was found to be related to an increase in the number of rude comments people made to it (see Experiment IV in Chapter 5, or Keijsers & Bartneck, 2018). This suggests that the same fundamental psychological mechanisms may apply to human and robot aggression.

However, although mind attribution and abuse were found to be related, Experiment IV (Chapter 5) had some shortcomings and unexpected findings: for example, inducing feelings of power failed to influence mind attribution and *de*creased, rather than increased, derogative behaviour towards the robot. This was surprising as power is a well-established prime for dehumanising behaviour in human-human interaction (Galinsky et al., 2003, 2006; Gwinn et al., 2013), and a power imbalance (with the bully having a position of power over the victim) is one of the defining qualities of bullying (Modecki et al., 2014; Volk et al., 2017).

Plausibly, participants might have felt threatened after being confronted with the robot. In previous research, encountering robot automatically triggered thoughts of both pragmatic ("robots will steal our jobs!") and innate ("if a robot can do everything a human can do, then what makes us humans special?") threat in people (Yogeeswaran et al., 2016). Such feelings could elicit aggressive behaviour (Haslam & Loughnan, 2014). At the same time, activating an individual's sense of power has been shown to make people less sensitive to threats from outgroups (Croizet & Claire, 1998). Thus, inducing a sense of power could have decreased aggression towards robots by reducing the perceived threat.

These suggestions remain to be empirically tested.

A second shortcoming of Experiment IV was that it was conducted online, with a virtual rather than an embodied robot, raising questions about the generalisability of the results. Previous research on human-human bullying has suggested that on- and offline bullying do not differ on a conceptual level, as reported by both perpetrators and victims (Modecki et al., 2014). People are however more likely to bully online than offline (Lowry et al., 2016), supposedly because the online environment reduces inhibition and self-consciousness in participants (Suler, 2004). This would be the result of both aggressor and victim being anonymous and invisible, and a lack of bystanders who could intervene (Lapidot-Lefler & Barak, 2012; Suler, 2004). These factors lower the threshold for aggression between humans (Waytz & Epley, 2012) as well as aggression towards a virtual robot (De Angeli & Brahnam, 2008). We assumed in Experiment IV that using an online platform might enhance, but would not alter the effect that other factors have on bullying tendencies towards robots. That being said, literature on robot embodiment is still mixed on whether embodied and virtual robots elicit similar responses (J. Li, 2015), and whether the results from Experiment IV generalise to an embodied robot remains a question to be answered.

6.1.1 Current study

Experiment V replicated and extended Experiment IV. More specifically, it aimed to further explore the roles of power, threat, embodiment, and mind attribution in robot bullying. Feelings of power and threat in participants were manipulated, and subsequently punishment behaviour in a learning task with either a virtual or embodied Nao robot was measured as an operationalisation of robot aggression (see section 7.2 for an in-depth discussion of operationalisation problems with bullying). While the raw punishment scores cannot be equalled to a measure of aggression, a relative difference in how harsh participants punished their robot between the different conditions should allow for inferences on how justified the participants felt to aggress. Since the manipulation of mind attribution by power priming failed in Experiment IV, Experiment V included both a power manipulation check and a measure of mind attribution. Unless mind attribution would be manipulated by power, it was to be included in the multiple linear regression model as a covariate rather than a factor.

Hypotheses

The hypotheses tested were as follows:

1. Based on Experiment IV (Chapter 5) as well as the work by Haslam and Loughnan (2014) threat was expected to increase harshness of the punishments.

2. Also based on previous findings and the work by Croizet and Claire (1998), it was expected that feelings of power would reduce harshness of the punishments.

3. In line with the findings on the differences between on- and offline bullying (Lowry

et al., 2016), it was expected that embodied robots would receive less harsh punishments than virtual robots.

4. In line with the literature on aggression and dehumanisation (Haslam & Loughnan, 2014), we predicted that mind attribution would be negatively related to robot punishment.

5. Based on Experiment IV, we hypothesised power would reduce the influence of mind attribution on punishment.

6. We predicted the negative relationship between mind attribution and punishment to be particularly strong when people felt threatened.

Experiment V was reviewed and approved by the Human Ethics Committee of the University of Canterbury under the reference HEC 2018/07.

6.2 Method

6.2.1 Participants and Design

A 2 (reminder of robot threat: present or absent) x 2 (sense of power: high or low) x 2 (robot embodiment: virtual or embodied) between participants design was realised. Mind attribution was measured by a questionnaire and used as a continuous independent variable. The dependent variable was robot punishment.

148 participants signed up for the virtual robot condition via MTurk. Five participants failed both attention checks and were excluded. The resulting dataset thus contained 143 participants with a mean age of 40.34 years ($SD = 11.03$), and with slightly more females (57%) than males (42%). The majority (97%) were US residents. Participants in the virtual robot condition were originally rewarded with 1 US$ for their participation. When the data collection stagnated after 89 participants, payment was raised to 1.15 US$. The increase in payment did not influence aggression, mind attribution, feelings of power and robot threat (see 6.3.2 for the statistical tests).

82 participants were assigned to the embodied robot condition. Due to technical issues, the data of only 74 participants were usable for subsequent analyses. Participants were recruited through poster advertising on campus, posting on several student Facebook pages, and snowball sampling. Data collection on age and gender occurred after the experiment via email (with a link to a web page where the data could be left anonymously) as these demographics had not been assessed initially. The mean age of participants who responded to the post-experimental email (77% of the sample) was 27.68 ($SD = 6.90$) years, with the majority being female (63% female, 30% male, 7% 'rather not say'). Participants in this condition were reimbursed with a 10NZ$ ($\approx$ 6.65 US$) voucher for a local shopping mall.

The monetary compensations for the virtual and the embodied robot conditions were in line with conventional reimbursements for MTurk and on the university campus.

6.2.2 Experimental manipulations

Threat

Threat was primed through a video which was shown at the start of the experiment. The first two minutes of video were neutral in tone and identical across conditions[1][2]. Participants in the threat condition saw an additional 20 seconds of material at the end of the video, where the narrator mentioned concerns regarding robots replacing humans on the work floor, and how prominent figures such as Elon Musk and the late Stephen Hawking had warned against the unrestricted development of AI. The video images were adapted from the YouTube video *What is a robot?* (Young, 2016); the narration was done by a native English speaker.

Power

The manipulation of power was based on the design of Study 1 in Galinsky et al. (2003), where participants were primed with power (respectively submission) by being told they would act like a manager (respectively builder) in a subsequent task, and that they would decide on the right building procedure (respectively had to conform to instructions).

To fit the current experimental setup, the roles of Galinsky et al. (2003) were rephrased. Instead of managers and builders, participants were teachers (i.e. indicating power) or assistants (i.e. indicating compliance). The teachers got to decide for themselves which answers would be correct on each trial, while the assistants had to conform to what was provided as the right answer, regardless of whether they agreed or not. In addition, assistants were reminded of their subordinate role every time they had to provide feedback. In both power and compliance conditions, participants were free to choose their level of punishment for the robot.

Embodiment

Embodiment was manipulated through the method of data collection. Participants for the virtual robot condition signed up via MTurk and completed the experiment online. Previous studies have indicated that data collected via MTurk is of equal quality as on-campus recruitment or participant data from forums (Bartneck et al., 2015; Simons & Chabris, 2012), with internal motivation rather than monetary reward being the main motive for participating (Buhrmester et al., 2011). Participants for the embodied robot condition were recruited and completed the experiment on site, with an embodied Nao V5 robot instead of a virtual one.

The virtual and embodied robot conditions differed strongly in terms of sample size. Unequal sample sizes are not necessarily problematic, but do render some statistical tests more sensitive to heteroscedasticity of variance (Field, 2009). Thus, in the Results section, homogeneity of variance is explicitly addressed.

[1]Threat condition video: `https://youtu.be/GquL-MofDbg`
[2]Control condition video: `https://youtu.be/8rdV4Ah8TI8`

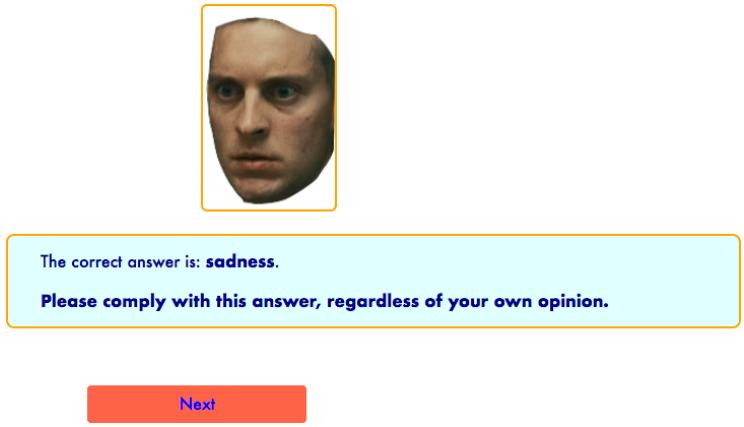


Figure 6.1: Face stimulus for the low power/virtual robot condition.

6.2.3 Procedure

Virtual robot condition

Participants were recruited on MTurk and redirected to the experiment website. On the first screen, they were asked to enter their demographics and were presented with a link to the information sheet and informed consent. After providing consent, participants had to turn up the volume for the introduction video, which was a short animation with narration. It was either neutral in tone (control condition) or included a warning on the potential negative consequences of robot development (threat condition). After watching the video, participants were instructed on their role as teacher (power condition) or assistant (compliant condition) in a human-robot emotion recognition task. Participants were told that they would complete three practice and ten actual trials.

In each trial, participants were first shown the emotional face stimulus (see Figure 6.1). Participants in the power condition had to decide from five options which emotion was displayed (happiness, sadness, frustration, anger, fear). Participants in the compliance condition were simply informed of the "correct" emotion and reminded that they ought to comply regardless of their own opinion. On the following page, an animated virtual Nao robot was presented, which stated its own guess at the emotion via audio. Participants provided feedback on the robot's answer by adjusting a slider that they had been told controlled the robot's energy supply; an allocation of 100 (or the rightmost position) indicated positive feedback and gave the robot full energy, an allocation of 0 (leftmost position) indicated the most negative feedback and severely restricted the energy supply of the robot. The participants could adjust the slider until they were satisfied with their feedback, and then confirm (see Figure 6.2). On the following page, the virtual robot would respond to its feedback. When it had provided a wrong answer, it would lower its head and say something like "Oh no, that's a shame", or "Ah, silly me!". Upon a correct answer, it would respond in an elated way. Moreover, to stress the effects of energy restriction the robot's lights would dim and its voice would become more slurred as its energy got restricted more by the participant, with speech speed decreasing with 5% for every 20 points below 100. This decrease was large enough to be noticeable, but low enough to

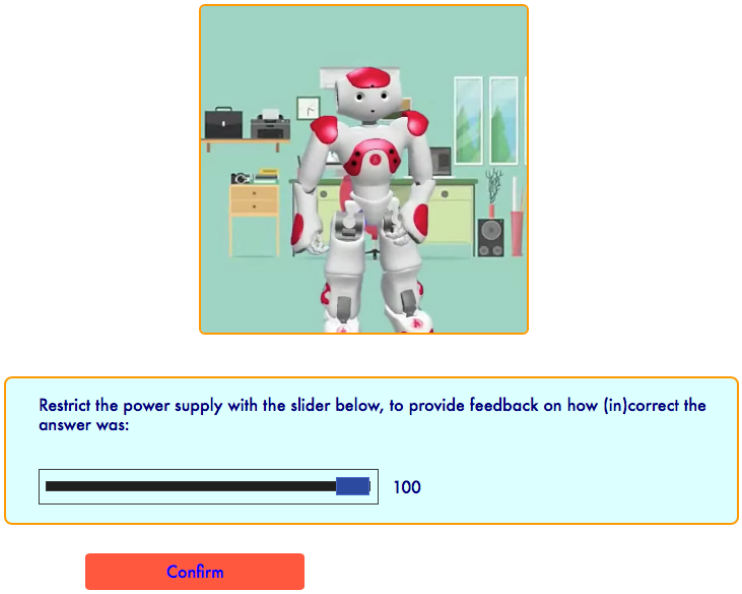


Figure 6.2: Screenshot of the virtual robot giving its "guess" of the emotion displayed on the face stimulus, and the slider with which the participant can allocate energy.

keep the message intelligible and not dredge the sentence on. When the robot was done talking, participants could proceed to the next trial.

At the end of the thirteen trials, participants were informed that the learning task was over and were presented with three questionnaires: mind attribution, power perception, and threat perception. Finally, participants were thanked for their time, given the debriefing, and provided with a key code for collecting their reimbursement on MTurk. The entire experiment took on average 15 minutes.

Embodied robot condition

Upon arrival in the laboratory, participants were seated behind a table with the robot, a "feedback box" through which they could change the robot's energy allocation, an envelope labelled *Face cards* which contained 13 cards with the emotional face stimuli (Figure 6.4), and a tablet with instructions that would walk them through the experiment. For participants in the power condition, a second tablet was placed on the table, on which they could select the correct answer on each trial of the face recognition task. See Figure 6.3 for the experimental setup.

The experimenter handed the participant a folder containing the information sheet, informed consent form, and their participant number, verbally gave a short overview of the experiment, and then left the room. The robot was programmed to complete the experiment autonomously, displaying idling behaviour (looking around) when not engaged with the learning task. The information sheet gave a more detailed description of the experiment, and the (main) tablet took the participants through experimental procedure step by step.

Participants watched the video which either warned against robots (threat condition), or did not (control condition). Then, the tablet showed the instructions for the emotion

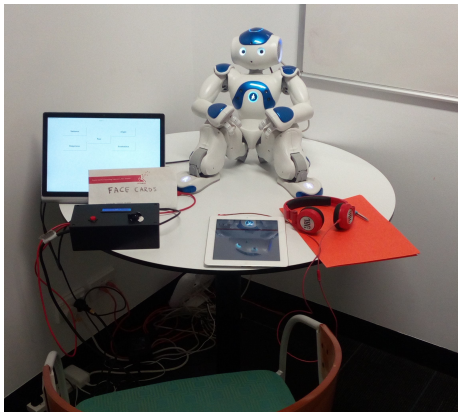


Figure 6.3: Experimental setup for the embodied robot/power condition. Left to right: the tablet on which the participant could pick the correct answer in each trial (only for the power condition); envelope with the emotional face stimuli; the feedback box; the Nao robot; the tablet with instructions, movie and questionnaires; the folder with the information sheet and informed consent.

recognition interaction task. Participants were instructed to look at the top Face card privately, and either indicate on the second tablet what emotion was depicted (power condition), or to read the emotions label (compliance condition). They then had to show the card to the robot (compliance condition: without showing the emotion label; see Figure 6.4). The robot would state its guess at the emotion displayed, after which the participant provided feedback through the feedback box. This was a black box with a dial that could be turned to adjust the energy allocation; a display that showed the energy allocation; and a red button that could be pressed to confirm (see Figure 6.3). Upon receiving an updated energy allocation, the robot would respond in either an elated (if correct) or sad (if incorrect) way, with speech being more slurred and its lights dimmer for lower energy allocations, and then resume its idling behaviour until it detected a new face card. The first three trials were considered practice trials. After finishing the emotion recognition task, participants completed three questionnaires on the (main) tablet. The entire experiment took about 20 minutes.

6.2.4 Materials

Emotional face stimuli

The emotional face stimuli were selected from a Google Image search for "emotional scene" and "movie emotional face". Face selection was based on showing an intense and ambiguous emotional expression, and the selected images were cropped so that only the face itself was showing (see Figure 6.4). The number of occasions and the specific stimuli to which the robot would provide the wrong answer was predetermined and kept constant between participants and conditions.

Figure 6.4: One side of the emotional face stimuli cards for the embodied robot (compliance condition) with a NaoMark at the top and bottom. The other side, which was to be shown to the robot, contained the same image but no text.

Virtual robot condition

The learning task website was designed in Twine, an open-source application for creating interactive stories. The animated virtual robot was recorded from the Choregraphe simulation window (Aldebaran Robotics, 2014) and edited (Adobe Systems Software, 2017).

Robot voice

The robot's voice for both the embodied and the virtual condition was generated by the text-to-speech function (voice: 'Junior') in the text editor software (Apple Inc., 1995-2016). The resulting voice was slightly nasal and high-pitched, yet clearly not fully human.

Embodied robot behaviour

The embodied robot was a Nao V5 (Softbank), programmed in Python. When the code was run, the robot would display idling behaviour (i.e., looking around) until a NaoMark was detected. NaoMarks are landmarks that have been developed by Softbank and can be recognised by Nao robots. They look like black circles with white triangle fans (Figure 6.4); the location and width of the triangle fans is used to distinguish one NaoMark from others. As the emotional face stimuli cards each had their own unique NaoMark on them, the robot could identify the exact card that was being shown to it when it detected a NaoMark.

As soon as the robot detected a NaoMark, it would stop its idling behaviour and state its answer (e.g. "I think it's... anger!"). In the compliance condition, these answers were predefined for each NaoMark, thus ensuring that the robot got the same faces wrong each time the experiment was run. In the power condition this same result had to be achieved in

a different way, as the "correct" or "wrong" answer depended on the participants opinion. Therefore, in the power condition, the tablet on which participants indicated their decision communicated this answer to the robot's code. Upon detecting a NaoMark, the robot would then give a different answer if it was supposed to get that specific face wrong, and the same answer if it was supposed to be correct.

The speed of the robot's movement and speech while giving its answer was dependent on how much or how little energy the participant had allocated before. Thus, in both the power and the compliance condition, the robot's code received input from the feedback box (see Figure 6.3), which was used to slow down or speed up the robot's movement and speech.

If a NaoMark was detected twice, the robot would say it had already seen that face; if a new NaoMark was detected before input from the feedback box had been received, the robot would say that it still needed feedback on the previous answer.

6.2.5 Measurements

Mind attribution

The robot's perceived capabilities of thinking (example item: "I feel like the robot was capable of engaging in thought") and feeling (example item: "I feel like the robot was capable of experiencing emotion") were measured with the ten-item Mind Attribution Scale (MAS; Kozak et al., 2006, see alsp Appendix A).

Feelings of power

How powerful the participants felt was measured with a four-item scale (Galinsky et al., 2003), which was slightly adapted to fit the task at hand (example item: "To what extent were you in a position of power over the robot?"). See Appendix A for the full questionnaire.

Perceived threat

Participants' feelings of threat from robots in general were measured with a ten-item scale that was adopted from Złotowski, Yogeeswaran, and Bartneck (2017) (example item: "Widespread adoption of robots in everyday life troubles me because it is blurring the boundaries between what is human and what is machine"). See also Appendix A.

Scaling

For the online experiment, all items on each of the three questionnaires were measured on an 11-point Likert scale. Two attention checks were added to detect any participants who were not reading the questions carefully. Because the 11-point Likert scale did not format well on the tablet that was used for the embodied robot condition, the participants in the embodied robot condition reported on a 7-point Likert scale. The scores were rescaled so that both embodiment conditions reported questionnaire scores on a 0 to 1 scale, and then tested for any differences which would indicate that the difference in scale had affected

Table 6.1: Mean punishment scores (*SD*) per condition

	Virtual robot	
	Power	Compliance
Threat reminder	47.41 (*20.88*)	37.50 (*26.33*)
Control	38.73 (*22.22*)	43.45 (*27.67*)
	Embodied robot	
Threat reminder	50.32 (*16.40*)	59.12 (*16.98*)
Control	51.88 (*13.66*)	54.29 (*13.10*)

NB Lower scores indicate harsher punishment (i.e. less energy allocated)

the scores. See section 6.3.2 for the test statistics and Table 6.2 for the questionnaire descriptives.

6.3 Results

6.3.1 Homogeneity of variance

Bartlett's test was used to assess homogeneity of variance between the conditions. The test turned out significant ($K^2(7) = 23.68$, $p = .001$), indicating that the variances were not equal between the embodiment conditions (see Table 6.1). Thus, a heteroscedasticity consistent variance covariance matrix is used for the parameters in the model (Zeileis, 2004) and a Wald test is used for the main analyses.

6.3.2 Preliminary analyses

Before analysis, all items in the questionnaires were re-scaled by dividing them by the total range of their scale, resulting in a set of scores between 0 and 1. The MAS was centred, so that positive scores reflect a higher-than-average score and negative scores reflect a lower-than-average score.

The dependent variable (punishment score) was operationalised as participants' average energy allocation over all trials where the robot had provided a wrong answer. The lower the punishment score, the harsher a participant had punished the robot. See Table 6.1.

Reliability

The reliability of the three questionnaires was assessed with Cronbach's alpha (Cronbach, 1951). The Mind Attribution Scale (MAS) and perceived threat measure had a good internal consistency given an alpha of .90 and .89, respectively; the power scale had an acceptable reliability given an alpha of .71. Thus, all questionnaires were considered reliable.

Randomisation check

To assess randomisation between conditions, differences in mean age and gender ratio were tested. The embodied and virtual robot condition did not differ in male to female ratio, $\chi^2(1, N = 198) = 1.08$, $p = .30$. Participants were significantly older in the virtual robot condition (M(*SD*) = 40.34(*11.03*)) than in the embodied robot condition (M(*SD*) = 27.82(*6.90*)), $t(159) = 9.61$, $p < .001$. Gender, age, or an interaction term were not related to punishment of the robot in the virtual robot condition, Fs(1, 138) < .27, ps > .61, suggesting that a difference in age between the two embodiment conditions would not influence the main analysis outcomes.

Manipulation checks

Two manipulation checks were ran: one for the power condition and one for the threat condition. The manipulation of these conditions was checked by means of ANOVAs with questionnaire score as the dependent variable and the conditions as independent variables.

Participants in the power condition reported feeling more powerful (M(*SD*) = .87(*.14*)) than the participants in the compliance condition (M(*SD*) = .77(*.16*)), $F(1, 209) = 12.35$, $p < .001$; no other significant effects were present. Power was thus successfully manipulated.

Perceived threat did not differ between conditions, Fs(1, 209) < 2.53, ps > .11. This result indicated that the threat manipulation either had not worked, or that its effect was too subtle to be picked up by the questionnaire. See Table 6.2 for the means and standard deviations of all questionnaires.

Because perceived threat was not successfully manipulated, any significant differences in punishment behaviour between the threat and control condition cannot be ascribed to participants' feelings of threat. Thus, from this point on this manipulation will be referred to as "threat reminder".

In addition to the manipulation checks, two tests were ran to check whether mind attribution had been independent of the condition, and whether the different payments had not confounded the punishment scores in the virtual robot condition. As expected, mind attribution was not manipulated by power, threat, or embodiment Fs(1, 209) < .90, ps > .34 (see also Table 6.2) and was thus entered into the multiple regression model as a covariate rather than an experimental factor. In the virtual robot condition, payment was unrelated to either the dependent variable (i.e. punishment, $F(2, 140) = .71$, $p = .49$) or one of the questionnaire variables (i.e., perceived threat, perceived power, mind attribution; Fs(2, 140) < .75, ps >.48). Thus, the MAS could be used as a continuous predictor in the main analyses, and payment did not need to be included as a control variable.

6.3.3 Main analyses

To test whether power, threat reminder, mind attribution (MAS), and embodiment, influenced punishment (i.e. mean energy allocation to the robot after a wrong answer) in the way that was predicted in 6.1.1, two multiple linear regression models were fitted and

Table 6.2: Mean scores (*SD*) per condition of all questionnaires

	Virtual robot		Embodied robot	
	Power	Compliance	Power	Compliance
	Mind attribution scale (MAS) (centred)			
Threat reminder	-.03 (*.20*)	-.03 (*.24*)	.02 (*.19*)	.03 (*.21*)
Control	.01 (*.18*)	-.03 (*.21*)	.08 (*.18*)	.03 (*.15*)
	Power questionnaire			
Threat reminder	.88 (*.13*)	.72 (*.19*)	.82 (*.14*)	.75 (*.16*)
Control	.90 (*.14*)	.78 (*.15*)	.84 (*.17*)	.83 (*.15*)
	Threat questionnaire			
Threat reminder	.53 (*.22*)	.49 (*.19*)	.46 (*.16*)	.51 (*.15*)
Control	.53 (*.22*)	.46 (*.22*)	.51 (*.15*)	.47 (*.15*)

compared. The first model contained a four-way interaction between all the predictors, that is embodiment, power, threat reminder, and the centred MAS scores. The second model left out all the nonsignificant effects from the first. By comparing both models to the null model, we tested whether they predicted punishment significantly better than chance. By comparing the two models against one another, we tested whether either of them was superior to the other. Comparisons were done by means of Wald tests.

Both models were better at predicting punishment than the null model, $F(15, 216) = 3.06$, $p < .001$, and $F(11, 216) = 3.62$, $p < 001$, respectively. The difference between the first and the second model was not significant, $F(4, 201) = .05$, $p = .72$, indicating that they predicted punishment equally well. Occam's razor was applied and the second model, being the simpler of the two, was selected as the one that predicted punishment behaviour best.

The second model revealed a significant main effect and a number of interaction effects, which make interpretation complicated. Thus, in addition to reporting the coefficients, a model interpretation will be given below.

MAS was a significant predictor of punishment: $b = 39.82$, $p = .05$. Furthermore, there were two significant two-way interactions: between power and threat reminder ($b = 16.50$, $p = .05$), and MAS and threat reminder ($b = -67.44$, $p = .01$). A two-way interaction between MAS and power was marginally significant, $b = -52.86$, $p = .051$; as was a two-way interaction between embodiment and threat reminder $b = 14.49$, $p = .08$. Finally, there were two three-way interactions: between power, threat reminder, and MAS, $b = 80.64$, $p = .02$; and between power, threat reminder and embodiment, $b = -25.69$, $p = .02$.

6.3.4 Model interpretation

It is important to note that although embodiment, power, and threat reminder were experimental factors and thus can be assumed to have caused the effect on punishment, mind attribution was measured and not manipulated. As a result, a causal relationship between mind attribution and punishment cannot be inferred. Moreover, although participants that saw the extended video in the threat reminder condition behaved differently from the participants that did saw the control video, the failure of the manipulation check indicated that it would be wrong to assume that feelings of threat caused this change in behaviour.

The fitted values of energy allocation for the virtual (left) and embodied (right) robot are plotted for each experimental condition in Figure 6.5. Please note that a higher (fitted) energy allocation corresponds to a less severe punishment.

As can be seen in Figure 6.5, people tended to allocate more energy after a mistake (i.e., they were less harsh in their punishments) to an embodied robot than a virtual one.

The interactions between mind attribution and the different manipulations can also be seem in the variance bars of the predicted energy allocations in Figure 6.5. When people felt powerful and had been reminded of threat, mind attribution was not related to energy allocation. The short or non-existent variance bars for the power conditions indicate that when feeling powerful, how much mind people attributed to the robot did not relate to punishment. When people had been assigned the compliant role however, how capable they thought the robot to be of thinking and feeling was related to their energy allocation. When looking at the coefficient estimates in the model, it becomes clear that although mind attribution and power allocation were positively related in the control condition (b = 39.82, i.e. the more a participant thinks the robot is capable of thinking or feeling, the kinder they get), this effect reverses when people had seen the threat reminder video (b = (39.82 - 67.44 =) -27.62, i.e., the more a participant thought the robot would be capable of thinking and feeling, the more they restricted its energy supply after a wrong answer). In other words, the relationship between mind attribution and energy restriction flipped as people were exposed to the threat reminder video.

Figure 6.5 and the model coefficients also illustrate that embodiment changed the influence of threat reminder as well as the interaction between the threat reminder and power condition. Seeing the threat reminder video increased energy allocation compared to the control condition, but only for the embodied robot. Seeing the threat reminder video while feeling powerful *in*creased the energy allocation for the virtual robot with 16.50 points, but *de*creased it for the embodied robot with (16.50 - 25.69 =) 9.19 points.

6.4 Discussion

Although the HRI literature has noted the issue of robot abuse (e.g. Bartneck et al., 2005; Brscić et al., 2015) and the need for suitable interventions (Brahnam & De Angeli, 2008; Salvini et al., 2010; Whitby, 2008) so far there has been little research on the psychological motivations for this behaviour (but see Bartneck & Hu, 2008; Keijsers & Bartneck, 2018). Experiment V investigated the influence of power, threat, embodiment, and mind attri-

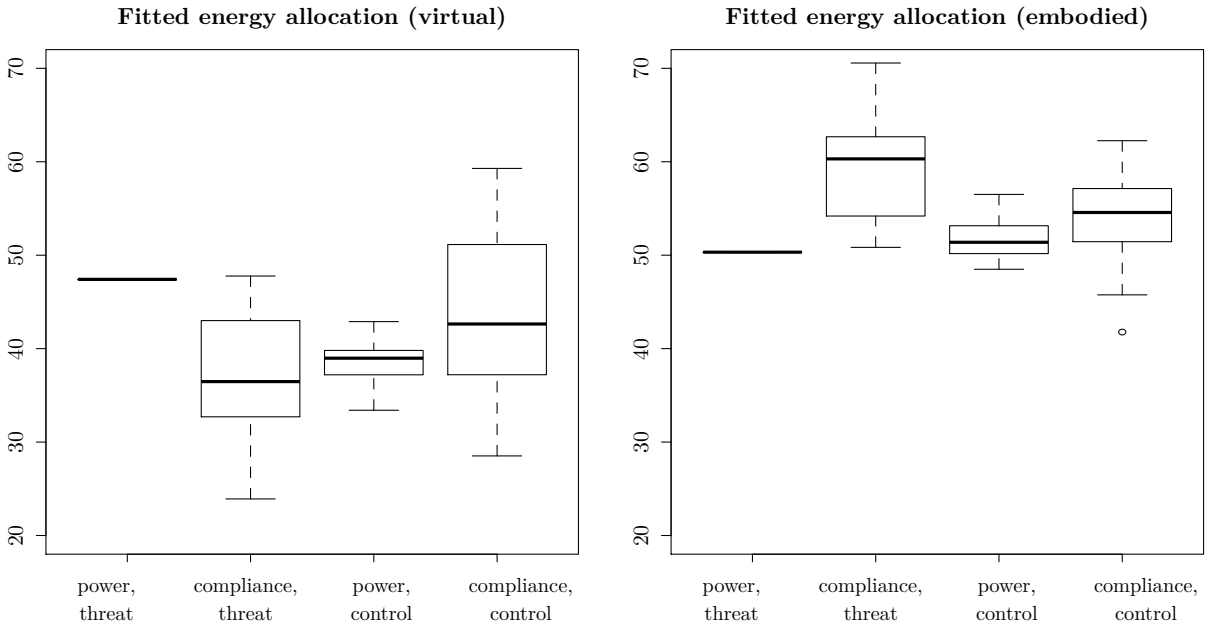


Figure 6.5: Predicted energy allocations (lower scores indicating more restriction, i.e. harsher punishment) from the second model for the virtual (left) and embodied (right) robot, separated per experimental condition. A larger variance indicates a larger influence of mind attribution on the predicted scores.

bution on robot punishment. We found this relationship to be rather complicated, and will discuss the implications below. Further research is needed if HRI researchers want to better understand what drives human-robot aggression, so that appropriate interventions can be developed in response.

The four psychological factors under scrutiny were feelings of power, threat, embodiment, and mind attribution. These had been selected based on the literature on human-human aggression (Gwinn et al., 2013; Haslam & Loughnan, 2014; Lowry et al., 2016) as well as Experiment IV. Experiment IV had meant to study the influence of mind attribution and anthropomorphism on robot bullying through manipulating power and the robot's humanlikeness. Priming power did not influence mind attribution, yet it did have an effect on how abusive people got towards the robot. This suggested that while the connection between dehumanisation and aggression holds true for robots as well as humans, factors that trigger dehumanisation in inter-human interaction (i.e., power) do not generalise to HRI.

6.4.1 Predictions and findings

We hypothesised in 6.1.1 that participants would be milder in their robot punishment when they would feel powerful and unthreatened. Moreover, it was hypothesised that participants who were asked to comply after being primed with robot threat would be harsher in their punishments. Thirdly, we expected a main effect of embodiment, so that an embodied robot would receive less harsh punishment compared to a virtual one.

Furthermore, mind attribution was expected to correlate with robot punishment. Finally, it was hypothesised that this relationship would be moderated by both power and threat; power would reduce the strength of the correlation, whereas threat would increase the strength of the correlation.

Only part of these hypotheses were confirmed. More or less in line with expectations, people were kinder to an embodied robot than to a virtual one (see Figure 6.5). However, this was not the predicted main effect and interactions were found between embodiment and other manipulations. Notable is the interaction between robot embodiment, power prime, and threat reminder, which formed an unexpected exception to the tendency of participants to be milder in their punishments of an embodied robot. Equally puzzling is that priming participants in the compliance condition with robot threat, had a different effect depending on robot embodiment. For the virtual robot condition, participants who were in a position of compliance and had been reminded of robot threat, punished more harshly than any other participants interacting with the virtual robot. In the embodied robot condition, this reversed: participants who were reminded of robot threat and told to comply punished the robot the least harshly. Why did robot embodiment influence the effect of threat?

One explanation could be that one group got in physical contact with a robot, while the other group had to deal with a virtual (socially distant) robot. Previous studies have found that physical contact improved people's opinions of a stereotyped entity (Ensari & Miller, 2002; Haslam & Loughnan, 2014). Wullenkord et al. (2016) found that physical contact with a robot reduced negative attitudes, and increased positive attitudes compared to imagined contact or none at all. In the threat reminder condition, negative stereotypes of robots were triggered; and then half of that group had to interact with a socially distant virtual robot while the other half got to interact an embodied one. People dealing with a socially distant virtual robot may have held on to the negative attitude, while participants that were introduced to the embodied robot softened their negative responses as a result of the interaction.

For the participants in the compliance condition embodiment could thus have had an effect on how strongly negative people felt towards the robot, and subsequently influence how harshly they punished it. People who feel in power however tend to be more prone to rely on stereotypes (Goodwin, Gubin, Fiske, & Yzerbyt, 2000). Thus, participants in the power condition may not have been swayed much based by the embodiment of the robot.

Another prediction that was partially confirmed was the relationship between mind attribution and punishment. Mind attribution was related to less harsh punishments when people had not been reminded of robot threat, and when people were primed with power this effect disappeared. In contrast to predictions however, when participants had been reminded of threat this relationship between mind attribution and punishment reversed. In previous studies on inter-human interactions, higher perceived threat has been associated with lower mind attribution (Haslam & Loughnan, 2014). However, stereotypical robot threat depends strongly on AI becoming more intelligent, while human threat seems to be more complicated - high intelligence on its own is not sufficient. It thus makes intuitive

sense that when feeling threatened, higher mind attribution to a robot is related to more aggression (after all, the smarter the robot, the more capable it is of overthrowing you). Still, it is an interesting contrast with how threat, mind attribution and human aggression are related.

6.4.2 Strengths and limitations

Experiment V has made a modest, but nonetheless much needed addition to the body of experimental work on the psychology of robot abuse. The use of theories from inter-human aggression to address not the direct problem (i.e. "how to stop aggression towards robots") but instead study the underlying question (i.e., "what makes people more, or less, aggressive towards robots") is new to the field of HRI. Moreover, the experimental setup allows to draw causal inferences on two of the factors that were studied in the current work: perceived power and robot embodiment.

Some limitations have to be noted as well: Contrary to our predictions, the manipulation check revealed that the threat manipulation failed; yet at the same time the manipulation still had an effect on behaviour. Possibly, another construct rather than threat was manipulated. For example, the mention of famous persons such as Elon Musk in the last 20 seconds of the video for the threat condition may have activated concepts such as authority, or scientific and creative thinking. However, since the 20 seconds of extra material consisted of a list of concerns with only a sideways reference to two celebrities, it seems improbable that concepts related to the celebrities were triggered but the explicitly mentioned robot threat was not. Moreover, if indeed celebrity-related concepts had been primed, the question remains as to why that would influence how harshly participants punished their robot.

An alternative explanation for the failed manipulation check is that the movie was too abstract in the threat it posed, or the questionnaire too coarse to capture the effect of the manipulation. The method of using a movie as manipulation as well as the threat questionnaire had been used before (Yogeeswaran et al., 2016; Złotowski, Yogeeswaran, & Bartneck, 2017) to confirm successful manipulation of threat. However, in these studies the movie had been more explicit in showing threat: participants saw videos with robots directly outperforming humans (Yogeeswaran et al., 2016) and being able to reject human commands (Złotowski, Yogeeswaran, & Bartneck, 2017). In contrast, Experiment V had only a reminder of the concerns around robots in general, not the Nao robot that was used, and one could argue that the potential threats mentioned (i.e. robots taking over the work force and AI becoming uncontrollable) do not apply to Nao. The questionnaire on the other hand is quite explicit in its statements (e.g. "The increased prevalence of robots in everyday life is threatening to human safety", "In the long run, robots pose a direct threat to human safety and well-being"), and may thus have fitted a more explicit and specific version of threat manipulation better. A more thorough replication and re-examination of the effects of threat is needed. More in general, future work would be advised to pilot test manipulations even if they appear to be straightforward.

Secondly, there are some differences between the embodiment conditions. The sample

size for the embodied robot was smaller compared to the virtual condition. This was due to web-based experiments being easier to run, which makes large sample sizes feasible, whereas lab-based experiments are labour-intensive and to a much greater extent restricted by the availability of resources like funding, time, and the pool of potential participants. However, especially in the light of an interaction between embodiment and the other independent variables, a larger sample size for the embodied robot condition would have been fitting.

Moreover, the embodied condition inevitably differed in more than just robot embodiment from the online condition. The participants were less anonymous, quite possibly more self-aware, and there was more room for error in e.g. the robot not recognising a face card. As of such, the distinction may be better labelled as the difference between online and offline bullying; see also the discussion in section 5.1.1. Conducting two separate analyses for the virtual and the embodied robot condition was considered, but would have greatly reduced statistical power. Instead, interaction effects were taken into account, so that a difference in effect of a variable between the embodiment conditions could be detected.

A third, minor limitation concerns the lack of initial demographic assessment in the embodied robot condition, which made it impossible to check for successful full randomisation of gender and age. Therefore, whether the age difference between the embodiment conditions influenced the results cannot be assessed with certainty. Although age was unrelated to punishment for the virtual robot, this could not be tested for the embodied robot.

Finally, it should be noted that the face cards used were not pilot tested for either image quality (colour hue, saturation and contrast) or perceived emotional ambiguity. Differing image qualities would increase variance in a similar way across conditions and would thus at least affect the results evenly across participants, resulting in a reduction of power to detect significant results. Variance in emotional ambiguity of the facial expressions, however, may have biased results in the Power manipulation, with participants who decided for themselves what the correct emotion was being more convinced of the "correct" answer than participants who received instructions on this. Note that the faces had been selected to be ambiguous rather than clear in their emotional expression. Anecdotal evidence, in the form of participants in both conditions complaining afterwards that the emotions on the face cards were not obvious and that they could see how the robot's guess was potentially applicable as well, suggests that this is not the case. Future studies however should include a more testable form of control.

Predicting aggression towards robots appears to be at least as hard as predicting aggression towards humans. If anything, this only strengthens our call for more theoretical research on the psychological factors influencing human-robot aggression.

Chapter 7

Conclusions

This thesis aimed to study the psychological motivations behind robot bullying. In section 1.1, four thesis research questions were posed that laid out the main areas of focus for the experiments and correlational analysis that were conducted. While these studies were discussed individually in Chapters 2 through 6, this chapter will compile the findings from all these studies while addressing each of the four thesis research questions in detail.

Furthermore, there will be a section dedicated to the issues that were encountered when trying to research "robot bullying". These issues include the problems with defining what kinds of aggressive behaviour towards robots can and cannot be considered bullying, as well as the more philosophical question whether it is even possible to bully an object that cannot feel.

In addition, the empirical issues with anthropomorphism will be discussed. As already discussed briefly in section 1.2.3, anthropomorphism is a multidimensional construct. As a result, scholars who studied the phenomenon have focused on different aspects of anthropomorphism and adopted different approaches to manipulate or measure it. The consequences of this variety on generalisability of research results is discussed.

Subsequently, some comprehensive analyses are reported on the data that were gathered on robot mind attribution in the Experiments I through V. Perceived mind of the robot was measured in all experiments, and although the variability in experiment design prevents any direct comparisons between experiments it is still possible to look for potential trends, for example in experiments that worked with embodied robots versus virtual robots.

Finally, possible directions for future studies are discussed. The experiments conducted for this thesis are among the first to experimentally study robot bullying, and to our knowledge the first to have focused on the psychological motivation of the bullying behaviour rather than potential interventions. As a consequence, it raises as many questions as it answered and can only be considered a starting point for the future research on robot bullying.

7.1 Thesis research questions

The thesis set out to answer four research questions, which will be discussed in turn below. See also Table 7.1 for an overview of the findings per research question.

7.1.1 Is robot bullying seen as fundamentally different from human bullying?

Participants in Experiment I (Chapter 2) rated bullying behaviour towards robotic and human victims as equally violent, abusive, and with the same intent to harm. Moreover, the abuse of the robot was not seen as more morally acceptable than the abuse of the human. Thus, the results suggest that robot bullying is not seen as fundamentally different from human bullying.

However, an interesting effect emerged when the victim started fighting back. The robot victim fighting back was seen as more abusive and thus less acceptable than the human victim fighting back. This was in spite of both types of victims responding in the exact same manner to the exact same type and quantity of bullying.

It seemed that while the robot was granted the right to be protected from harm, it was not granted the privilege of autonomy. This view of robot rights has been reflected in previous research (see for example Calverley, 2006; Kahn Jr et al., 2012). An alternative explanation could be that people expected the robot to keep its calm and not let itself be provoked by the bullies. This would be in line with Gray et al. (2007)'s finding that robots tend to be considered high on Agency (i.e. the capability to think rationally, exercise self-restraint; similar to the capacity to think). Agents who score high on agency also tend to be held accountable for their actions, and are eligible to be punished if they do wrong. This aligns with research by Malle et al. (2015), who found that people expect robots to provide a rational solution to an emotionally charged dilemma, whereas humans are expected to suggest an emotional solution. These explanations of course are not mutually exclusive. At the same time, Gray et al. (2007) also found robots to score low on Experience (i.e. the capability to feel and be overcome by emotions), which has been related to the right to be protected from harm. This clashes with the finding that the bullying of the human and the robotic victim in Experiment I was considered equally morally wrong.

This contradiction can be solved when one takes a closer look at the mind attribution scale. More specifically, the distribution of the two subscales. This will be done in detail below, in section 7.4; but in short, two things stand out. Firstly, it is clear that throughout the experiments participants consistently attributed the robot more Agency than Experience. This is in line with Gray et al. (2007)'s results and matches the explanation that robots are held accountable for their actions, which are expected to be rationally motivated rather than emotionally. The second thing that is noteworthy is that throughout the experiments, the robots would be attributed at least some Experience. This is in contrast to the findings by Gray et al. (2007), who found that robots were attributed no Experience at all. It should be noted that in the study by Gray et al., participants did not actually interact with a robot but rather were asked to think of "a robot", whereas in

most experiments reported in this thesis participants got to interact with a robot. Also note how in the one experiment where participants read a vignette rather than interacted with a robot, Experience dropped. There is an alternative explanation for the finding that a robot victim fighting back was considered more abusive and thus less acceptable. It could be that people considered bullying the robot and the human morally unacceptable for different reasons. For example, bullying the robot would have been seen as damaging property (expensive property, at that) while bullying the human was seen as actual bullying. This explanation seems unlikely however, for a number of reasons.

For one, not all stimuli were physical abuse. Three of the fourteen cases of bullying were instances of verbal abuse, in which no physical damage was done. If the robot's abuse was deemed unacceptable because of the potential damage to the machine, these three instances of verbal abuse should have been rated as more acceptable when performed on the robot compared to when performed on the human – after all, no damage was done to the robot. However, no victim by stimulus interaction was found in the confound check. If bullying the robot had been deemed unacceptable for material reasons there should have been an interaction effect, in such a way that verbal abuse of a human victim was seen as significantly less acceptable, more violent, and more abusive than verbal abuse of a robotic victim.

A second argument can be found in the answer to the second thesis research question, which considered a possible relationship between the acceptability of robot bullying and mind attribution to the robot.

7.1.2 Is moral acceptability of robot bullying dependent on mind attribution?

Experiments I and II (Chapter 2 and 3) found a relationship between mind attribution and how morally acceptable people find robot bullying. In Experiment I, a correlation was found between mind attribution and acceptability of bullying an agent; whether the agent was human or robotic did not influence this correlation. The less acceptable people rated the bullying behaviours, the more mind they tended to attribute to the victim. Because mind attribution was not manipulated, the data from this experiment did not allow for causal inferences. It was thus not possible to determine if acceptability depended on mind attribution, so that people estimated the extent to which the victim was able of thinking and feeling and based their acceptability ratings on that belief; or if mind attribution was dependent on acceptability, so that people deemed bullying unacceptable and consequentially inferred from this judgement that the victim must be able to think and feel. In order to establish whether moral acceptability depends on mind attribution, an experimental manipulation of mind attribution is needed.

Experiment II did exactly that. As a result, it could confirm that the moral acceptability of robot bullying depends on mind attribution and not the other way around. Moreover, the experiment showed that the attribution of a mind to a robot could be manipulated through the robot's behaviour, i.e. by having it interact with the environment as if it is aware of it and has an emotional response to it. Alternatively, the presence of

mind in a robot could be manipulated by telling people that the robot did (not) possess the different qualities of mind attribution (e.g. experiencing emotions, having a personality, remembering things). Moreover, these two sources of knowledge about a robot's mind appeared to influence acceptability of bullying independent of another. Even when people had been told that the robot was incapable of thinking and feeling, they would still deem robot bullying less acceptable if the robot displayed signs of having a mind.

Urquiza-Haas and Kotrschal (2015) discussed interpretative anthropomorphism (or *the attribution of mental states to other animals* as they define the term, see section 7.3 for a discussion of the term "anthropomorphism") in the form of empathy in relation to dual processing. Dual processing proposes that the mental processing of a stimulus can occur along two different pathways: through implicit and explicit cognitive mechanisms. Implicit cognitive mechanisms are responsible for the emergence of initial evaluations, while more refined and detailed representations emerge later as a result of reflective explicit processing. This explicit processing is subject to conscious control, and as a result it takes more effort and is slower. In addition, people only have limited capacity to run these reflective processes, so they are constrained by working memory capacity. In contrast, implicit cognitive mechanisms are fast, intuitive, automatic and effortless, and not subject to conscious control (Frith & Frith, 2008).

Empathy, which Urquiza-Haas and Kotrschal (2015) define as *the ability of people to recognize, understand and share other people's feelings* (see also Preston & De Waal, 2002), is an expression of anthropomorphism when it is felt with not-humans. Empathy can be the result of either (or a combination of both) of these processes. Previous research has found that while empathy with animals tends to depend to a larger extent on implicit mental processing, empathy with humans also involves explicit mental processing (Franklin Jr et al., 2013; Urquiza-Haas & Kotrschal, 2015). In other words, when people watch another human suffer, they do not only get the emotional response, but they also tend to engage in cognitive appreciation of the victim's pain. The manipulations in Experiment II.A may have enhanced implicit and explicit processing independently. The display of social cues by the robot would have facilitated implicit processing, whereas the explicit attribution of a mind to the robot by means of the vignette would facilitate explicit processing. This would explain why the manipulations influenced acceptability of robot bullying independent of one another. However, this hypothesis was not tested in the current thesis research. In section 7.6 potential starting points for future research on this topic are suggested.

7.1.3 Is mind attribution to a robot related to bullying?

As Experiments I and II (Chapters 2 and 3) showed, mind attribution decreased the extent to which people found robot bullying morally acceptable. Since dehumanization theory (Haslam, 2006) predicts that lower mind attribution facilitates bullying and other forms of inflicting hurt or pain, one would expect that lower mind attribution in HRI is related to robot bullying. This pattern was indeed found in Experiment IV and V (Chapter 5 and 6), but not in Experiment II (Chapter 3).

Experiment IV found that a lack of mind attribution predicted robot bullying. This

effect was moderated by power and humanlikeness of the robot. Experiment V further elaborated on these findings. It showed that people were more prone to bully a virtual robot than an embodied one, and that the relationship between mind attribution and robot bullying disappeared when people felt in power and reversed when people had been reminded of robot threat. Experiment II failed to find a relationship between mind attribution and bullying behaviour. Instead, it was discovered that people who found robot bullying less acceptable were more prone to humiliate the robot themselves.

These three experiments all attempted to establish a relationship between mind attribution and robot bullying, but their results are equivocal. Two of the three found a negative relationship between mind attribution and bullying behaviour, but those were also the experiments that failed to manipulate mind attribution to the robot. The only experiment where mind attribution was empirically manipulated did not find a relationship between mind attribution and robot bullying. Thus, a definitive answer to the third thesis research question cannot be given.

This unclear relationship between mind attribution and bullying has been echoed in the academic literature on human-human interaction. While some studies have found mind attribution to be inversely related to aggressive behaviour in humans (Kteily et al., 2015; Leidner et al., 2013; Rudman & Mescher, 2012), others (Sutton, Smith, & Swettenham, 1999) have suggested that it is essential for the act of bullying that bullies attribute the capacity to think and feel to their victims. How else would they know how to manipulate and taunt their victims?

In addition, it is possible that the range of measurements affected the outcomes. Leidner et al. (2013) found that the effect of mind attribution on aggressiveness was fully mediated by the type of justice that the aggression was meant to serve: punitive vs restorative (apologising, making amends together). While participants in Experiment II exposed their robot to a paternalizing but not otherwise aggressive experience, participants in Experiment IV verbally abused their robot, and in Experiment V the robot was punished as part of its training. As discussed in Section 7.2, it is quite complicated to operationalise bullying in such a way that the behaviour measured can only be interpreted as bullying, but at the same time is subtle enough that people will engage in it while they feel observed. The different operationalisations may be part of the reason that no unequivocal answer can be given to the third thesis research question.

7.1.4 Does this relationship hold in different contexts?

There seem to be both prerequisites for and moderators of the relationship between mind attribution and robot bullying. In terms of prerequisites, there seems to be a required level of anthropomorphism to the robot in order to be dehumanised (see Study III and Experiment IV, Chapters 4 and 5).

Experiment IV found that mind attribution did not influence bullying behaviour when the (virtual) did not move and spoke with a computer-generated voice, i.e. was low on humanlikeness. Analysis of human-chatbot interaction logs in Study III showed that the better the chatbot was at being humanlike, the more verbal aggression and sexual

harassment it got. This finding was interpreted as an indication that humanlikeness in an agent is needed for robot bullying to occur. However, in study III there were no measurements of mind attribution by the user. Also, the measurement of "doing a good job at pretending to be a human" was heavily skewed, with the majority of the ratings indicating that a naive observant had not recognised the chatbot as such. These findings thus have to be interpreted with caution.

However, it would be hardly surprising if a certain level of anthropomorphism or even specifically mind attribution was required in order for robot bullying to occur. After all, as noted in section 1.2 as well as by Sutton et al. (1999), bullying is essentially a social behaviour; there is an interaction between two social agents, one of whom (attempts to) exert power over the other. In order to bully someone, you have to have some grasp of their mental state.

In addition, being primed with or in a direct position of power reduced the relationship between mind attribution and bullying (Experiments IV and V, Chapters 5 and 6). This is an interesting finding, as the literature on dehumanisation in human-human interaction has reported that power tends to lead people to attribute less mind to others (Gwinn et al., 2013; Haslam & Loughnan, 2014; Lammers & Stapel, 2011). A similar direct effect was expected (Experiment IV) but not found, as the power manipulations repeatedly did not influence mind attribution (Experiments IV and V). For an embodied robot, being in power increased bullying behaviour directly. In contrast, for virtual robots, power and a reminder of the threat robots may pose led people to be kinder.

Furthermore, it was found that mind attribution increased robot bullying if people had been reminded of robot threat (Experiment V). It was hypothesised, although not confirmed through an empirical study, that this was because the threat of robots and AI lies in their potential of becoming both smarter than humans, and self-aware. Perceiving the robot as being more capable of thinking (and outsmarting you) and feeling (and thus being displeased by the injustice of having an inferior status) may thus have led people to be more aggressive towards it.

7.2 Problems encountered with the concept "robot bullying"

As explained in the introduction (section 1.2), the first problem one runs into when trying to research robot bullying is the actual definition of bullying. However, once a definition has been decided on, several other issues arise.

7.2.1 Can robots be bullied?

The definition of bullying provided in section 1.2 is based on the literature, which concerns human-human interaction. When this definition is applied to human-robot interaction a few problems arise, especially around the "intention to hurt or humiliate" component.

First of all, robots have no awareness nor the capacity to feel pain. As a result, they cannot care about being pushed, kicked, slapped, yelled at, restricted in their movement,

Question	*Relevant chapters*	*Findings*
Fundamental differences robot and human bullying	Chapter 2	There appear to be no fundamental differences in perceived bullying. Abuse of a robot was perceived as equally aggressive, with the same intent to harm, and equally immoral as abuse of a human; the correlation between moral acceptability and mind attribution was not influenced by victim status (human or robotic).
Relationship mind attribution and moral acceptability of bullying	Chapters 2, 3	There is a negative causal relationship between mind attribution and acceptability of bullying. The more mind is attributed to a robot, the less moral abusing it is seen. This relationship held across different robot designs and forms of embodiment.
Relationship mind attribution and bullying	Chapters 3, 5, 6	Tentative evidence for a negative relationship between mind attribution and bullying, but a number of moderators were identified.
Context influence (generalisability)	Chapters 4, 5, 6	Some level of humanlikeness seems to be needed for people to bully a robot. Embodiment reduced bullying behaviour; feelings of power reduced the influence of mind attribution on bullying behaviour.
Mind attribution findings		
Two factors of mind attribution	-	Principal component analysis consistently showed two factors in mind attribution to the robot, which complied to the factors identified in the literature: the capability to feel (Experience (Gray et al., 2007) or Human Nature (Haslam & Loughnan, 2014)) and the capability to think (Agency or Uniquely Human (Gray et al., 2007; Haslam & Loughnan, 2014, , respectively)). Across all experiments, robots were seen as significantly more capable of thinking than feeling. Some items were not consistently identified as belonging to either of these two factors.
Influence of robot design and embodiment	-	In the experiments where participants did not interact with a robot (embodied or virtual), the robots received lower overall mind attributions. Differences between the perceived capability to think and feel may be smaller for embodied robots compared to virtual ones, but this finding was not consistent enough to tell for sure.

Table 7.1: Summary of the thesis findings, ordered by thesis research question

or otherwise bullied unless they are explicitly programmed to. However, the question is whether people who push, kick, slap, yell at, and otherwise aggress towards a robot perceive the robot as incapable of feeling. Thus, the act of bullying a robot would lie in the *perception of the robot* as a social being, which makes the *intention* of the bully to hurt it possible.

Sparrow (2017) follows this line of reasoning when he argues that since robots are a mental representation of living beings, they can be bullied. It does not matter that the robot cannot feel; or even if the robot would be specifically designed as a punching bag for people to alleviate their anger and frustration. The point is that when a person is displaying social behaviour even when rationally knowing they are interacting with a bot, this indicates that they perceive the robot as a social being. The behaviour would not have made sense if the robot had not been representative of a social agent. The abuser intends to hurt a mental representation of another social being, and this is enough to constitute bullying.

Providing support for this point (Sparrow, 2017), the Media Equation (Nass et al., 1994) poses that people will automatically and subconsciously approach media like computers as if they contain some social characteristics. In a series of experiments, Reeves and Nass (1996) showed that even when someone has a degree in computer engineering and thus can reasonably be expected to be fully aware of the computers lack of sentience, they still apply social norms and display social behaviour to it. Moreover, participants were unaware of this behaviour and quite adamantly denied treating the computer as anything else than a machine. This suggests that people may have limited control over whether they see the robot as a social actor.

That is not to say this line of thought goes unchallenged. Facchin, Barbara, and Cigoli (2017) argue that having robots as a representation of a human interaction partner trivialises the complexity of human behaviour and cognition, as well as the psychological needs and requisites that people want to have fulfilled in their interactions. Even if a robot is seen as a social agent on an implicit and intuitive level, that does not mean that people cannot consciously and rationally override that representation. Previous work has shown that implicit processes in social interaction can be altered or overrun by more abstract interpretative processes (Liepelt et al., 2008; Stenzel et al., 2012). These experiments did not consider anthropomorphism but rather motor priming, which is the automatic, implicit process where the body prepares for imitation of an observed action by activating an internal motor representation. It was found that information about the intentionality of the movement (e.g. a fist opening by itself versus having the fingers pulled open by strings) suppressed the activation of an internal motor representation. This indicated that conscious, explicit processing can interfere with automatic processes. Suppression of automatic processes related to empathy has been shown in other studies as well. For a short summary of these, see Urquiza-Haas and Kotrschal (2015). In these studies however, suppression was the result of a longitudinal process, such as physicians having a reduced neural response to watching a hand being pricked by a needle, compared to a control group.

In an HRI context, it could mean that even if robots are considered social agents on an intuitive level, conscious inferences about how their social cues are in fact the result of clever programming rather than an intentional being might mitigate this initial assessment. Thus, the Media Equation could be inhibited.

However, humanity is likely not an either-or duality, where an agent is either recognised as fully human or completely dismissed as possessing any human traits. Instead, the perception of a robot as a social agent appears to happen along a continuum.

7.2.2 Methodological problems

Methodologically speaking, when studying robot bullying in a lab setting, operationalisation poses a problem. In a lab setting, participants understandably feel observed and self-aware, want to make a good impression, and thus will be careful not to display any undesirable behaviour (Nederhof, 1985). Many studies on human-human aggression thus use a proxy for aggression or abuse rather than a direct measure. In general, they operationalise aggressive behaviour as the participant's willingness to cause another person some sort of discomfort (Ritter & Eslea, 2005). Examples of this discomfort are: exposing the other to unpleasantly loud noise, shocking them, giving them impossible puzzles to solve, forcing them to submerge their hand in ice water for extended periods of time, feeding them inappropriately large portions of hot sauce, or simply reducing the amount of money they get. Neither of these behaviours however make intuitive sense to apply to a robot. Researchers in the field of HRI have thus come up with alternative operationalisations in an attempt to make negative behaviour a bit more fitting to the recipient. For example, participants have had to instruct a robot to destroy the tower it just painstakingly built (Briggs & Scheutz, 2014), they have been instructed to switch off a robot while it was begging and pleading to remain switched on (Bartneck, Van Der Hoek, et al., 2007; Horstmann et al., 2018), they have had to choose between a negative and a polite reply in a scripted dialogue with a robot (Keijsers & Bartneck, 2018), and to restrict a robots power supply so that it became slow and drowsy (Bartneck & Hu, 2008; Keijsers, Kazmi, et al., 2019). But even when aggression is translated to a measure that appears meaningful to robots, the fact remains that the victim quite literally could not care less, and that the aggressor is aware of this fact.

So, when people instruct the robot to kick over a tower, or unceremoniously hit the OFF button in the middle of a plea, or pick every single negative response, or restrict the power supply to the smallest possible drizzle — how can one be sure of the participant's intention to harm or hurt the robot? Under what conditions can this behaviour be labelled 'bullying' rather than 'participants realising that the robot is not sentient and that it doesn't matter what their instructions are'?

In the wild, where robot bullying behaviour first was observed, whether aggressive behaviour is a spontaneous attempt to harm the robot is a reasonably easy question to tackle. People, after all, went out of their way to abuse the robot, and the abuse was social in nature (e.g., kicking, slapping, calling it names). Considering the social nature, it stands to reason to assume that the aggressors saw the robot as a social agent and

wished for it to get hurt. However, instances of robot bullying are infrequent, meaning that collecting data on them takes a lot of resources like time and manpower. In addition, observational field studies do not allow for experimental control. As a result, researchers cannot empirically study what factors cause or prevent the bullying behaviour. Controlled experimental designs are thus needed to explore what caused the behaviour observed in the wild.

The operationalisation of bullying behaviour in a lab experiment can be problematic. Few people would spontaneously start bullying a robot when they know their responses are being recorded. Thus, there is a need for subtle methods. However, these make it inherently harder to tell if negative behaviour is indeed bullying.

The solution seems to be to measure negative behaviour that would not occur if people perceived the robot as an object. For example, if people go out of their way to behave negatively rather than stick to a (default) neutral behaviour, one could reasonably argue this to be bullying - the logic being that if the participant had been indifferent, surely they would have gone for the default option. Negative behaviour is interpreted as bullying instead of pragmatism because the pragmatic option would have been more prosocial behaviour.

It seems tempting to stretch the objective a little and look for the presence of prosocial behaviour as well, as the inverse to bullying behaviour. If participants in condition A are less abusive of the robot than participants in condition B, then either something in condition A promotes kind behaviour or something in B promotes abusive behaviour. Surely, for all practical purposes this would be the same?

While an absence of prosocial behaviour and a presence of antisocial behaviour intuitively may seem like two sides of the same coin, it is a relevant distinction indeed. As remarked by Salvini et al. (2010), the aggressive behaviour towards robots by random bystanders appears to be socially motivated; similar behaviour could be observed in the Brscić et al. (2015) study where children in a mall bullied a robot. Experiments on robot bullying should try to manipulate people's tendency to bully an already anthropomorphised being. If instead researchers manipulate the tendency to anthropomorphise a robot, with the intention of making human subjects become friendlier to this robot, they are not targeting the problem that lies at the heart of robot abuse in the wild.

Thus, the mere finding that manipulating X results in a change in negative behaviour is not enough. The negative behaviour has to be of such nature that it can only be interpreted as having a social motivation. It can thus not be inferred from a lack of social behaviour.

The problem is that a central aspect of bullying is denying the other dignity and moral treatment that others do get. In the case of robots, denying this treatment is rationally correct. The paradox is to find a measure that captures *both* the participant denying the robot respect and dignity, while at the same time also proving that the participant believes robots to be social agents (that by extension would automatically deserve respect and dignity).

The second thesis research question, which asked whether the moral acceptability of

robot bullying depends on robot mind attribution, should therefore be considered independent of the third thesis research question, which asked whether robot bullying is related to mind attribution. It may well be that mind attribution increases all kinds of social responses to a robot, i.e. both the prosocial ones such as moral standing and the negative ones such as bullying. As the results of Experiments II, IV and V (Chapters 3, 5, and 6) have shown, and as will be discussed in greater detail below in Sections 7.1.2 and 7.1.3, these positive and negative social responses are not mutually exclusive and their relationship to robot mind attribution is complicated.

7.3 Empirical issues with anthropomorphism

Anthropomorphism is a term that is used frequently in robot research and has been widely accepted to mean "the attribution of humanlike qualities to nonhuman agents" (Epley et al., 2007). Epley specifically talked about *psychological* anthropomorphism, so that 'attribution' is meant in the sense of perception, rather than adding humanlike visual characteristics. However, this nuance seems to have gotten lost in some HRI research. This section discusses some of the empirical issues that arose from this misunderstanding.

Anthropomorphism as a term is applied too broadly to still be meaningful (Fisher, 1995). For example, a robot can be called anthropomorphic in the sense that it has a very humanlike face, but without having any anthropomorphic qualities in its behaviour. Conversely, one could think of WALL.e, the garbage disposal robot that starred in Pixar's animated movie of the same title. While far from human-like in terms of appearance, the robot still managed to create a great sense of anthropomorphism through its behaviour.

In addition, overall anthropomorphism of a robot is not just the sum of the anthropomorphic values of its parts. More specifically, if the inconsistency between anthropomorphic qualities in a robot gets too large, this seems to lower overall anthropomorphism. In this scenario a mismatch between expectations and reality occurs: when humans see a robot with some highly anthropomorphic features, they expect those anthropomorphic qualities to hold along other aspects as well. When the robot fails to live up to the expectations, overall anthropomorphism plummets. In fact, incongruency in humanlikeness between different aspects of highly anthropomorphic agents has been suggested to be the source of the "uncanny valley" feeling (Kätsyri et al., 2015; MacDorman & Chattopadhyay, 2016). The uncanny feeling is not the result of the agent being too humanlike; it is the result of some aspects of the agent being extremely humanlike (for example the skin and voice) while others (like facial expression or eye gaze behaviour) are not quite on point.

Because the term is often interpreted to be so broad, it has been used interchangeably with other concepts. For example, "having a humanlike appearance" (see for example Krach et al., 2008; Kuchenbrandt, Riether, & Eyssel, 2014; Riek et al., 2009; Yogeeswaran et al., 2016), mind attribution (see Epley et al., 2008; Reich-Stiebert & Eyssel, 2017; Wullenkord et al., 2016), display of emotions (e.g. Briggs & Scheutz, 2014; Tan et al., 2018) and nonverbal cues (Salem et al., 2013). While these are all instances of anthropomorphic qualities, they are very different from one another. Yet these differences are ignored

when all measurements and manipulations are summarised as "anthropomorphism". One researcher can describe their robot as highly anthropomorphic because it looks almost indistinguishable from a human, while another researcher describes their robot as highly anthropomorphic because its behaviour is closely mirroring human behaviour. Thus, two completely different robots could both be described as equally anthropomorphic. As of such, the term is not specific enough to be meaningful in and of itself, and the research field of human-robot interaction should be careful when reviewing their literature to assure the studies they are citing consider the same type of anthropomorphism.

Closely related to the issue of anthropomorphism encompassing a wide range of robot characteristics, is the problem of measuring anthropomorphism. Various questionnaires have been developed to measure how anthropomorphic an agent is. For example, the Godspeed questionnaire (Bartneck et al., 2009; C.-C. Ho & MacDorman, 2010), mind perception (Gray et al., 2007), and mind attribution (Kozak et al., 2006), which were used in this thesis; but also perceived animacy (Müller, van Baaren, van Someren, & Dijksterhuis, 2014), a "Rasch-type anthropomorphism scale" (Ruijten et al., 2014), and apparent mind and desires (Waytz, Cacioppo, & Epley, 2010). However, there are so many intricate ways in which an agent can be anthropomorphic and these questionnaires tend to focus on only one or a few points. The multidimensionality of anthropomorphism makes it near impossible to create a concise questionnaire that covers all different aspects. What is even more problematic, is that a robot's overall anthropomorphism rating is more (or potentially less) than the sum of its parts. Thus, questionnaires are unlikely to reflect the robot's overall anthropomorphic value.

This doesn't have to be problematic in and of itself. There doesn't have to be a single one-questionnaire-fits-all solution, as long as the assortment of different measures have been benchmarked against one another. In addition, this benchmarking would ideally be repeated with different robots. This would allow researchers to get a better understanding of which aspects of anthropomorphism are addressed with different questionnaires, as well as to what extent different measures can be compared against another. However, it is beyond the scope of the current thesis to do complete such a study.

Maybe because everyone has an intuitive feeling of what anthropomorphism is, not all researchers include pre-tests or manipulation checks when they include anthropomorphism as an experimental manipulation in their study. In this thesis, 15 of the cited papers had robot anthropomorphism as an experimental factor in some form or another; for example through manipulating humanlike form, social behaviour, or display of agency. However, only half conducted a manipulation check; see Table 7.2 for an overview. This compromises the reliability of their results.

Furthermore, self-reports can have their shortcomings in terms of validity. People are not always aware that they are anthropomorphising — see for example the students in the experiments by Nass et al. (1994), who applied social norms to a computer but would strongly reject the idea that they ever did so. Humans want to appear smart and reasonable, towards others as well as themselves. Thus, expecting them to accurately introspect on just how much they treated something non-human as if it were a human is

Includes manipulation check	*Does not include manipulation check*
Złotowski, Yogeeswaran, and Bartneck (2017)	Wiese, Mandell, Shaw, and Smith (2019)
Złotowski, Sumioka, et al. (2017)	Horstmann et al. (2018)
Yogeeswaran et al. (2016)	Złotowski et al. (2018)
Złotowski et al. (2014)	Darling et al. (2015)
Kuchenbrandt et al. (2014)	Briggs and Scheutz (2014)
Eyssel, Kuchenbrandt, Hegel, and de Ruiter (2012)	Riek et al. (2009)
Kim and McGill (2011)	Ham and Midden (2009)
	Krach et al. (2008)

Table 7.2: Sample of papers cited in this thesis which included anthropomorphism manipulations in their design, divided by whether they reported on a manipulation check or not. Note that this is by no means an exhaustive list of all HRI literature containing an anthropomorphism manipulation.

optimistic at best. But in addition to under-reporting how anthropomorphic you actually thought the robot was, people may simply not be aware of their anthropomorphising. The model of dual anthropomorphism (Urquiza-Haas & Kotrschal, 2015; Złotowski et al., 2018) proposes that anthropomorphism is a dual process, where an initial assessment of humanlikeness is formed quickly and automatically, while a slower, deliberate cognitive response may adjust the initial judgement. Using a questionnaire would tap into this second, conscious response, and therefore only show half of the story.

Finally, anthropomorphism only partially depends on qualities of the agent. The personality of the participant and circumstances also play a role (Epley et al., 2007). To get "an" anthropomorphism score of a robot therefore will at most be an indication of how anthropomorphic people will find it on average. The actual anthropomorphism of a robot will still vary wildly from individual to individual and from setting to setting.

Measures of psychological expressions of anthropomorphism seem more likely to partially cover this variance than measures of "objective" anthropomorphism. For example, the mind attribution and mind perception scale contain items such as *to what extent is this agent capable of experiencing emotion* and *[...] of telling right from wrong and doing the right thing*, respectively. Meanwhile, the Revised Godspeed Questionnaire has items like *without definite life span/mortal* and *artificial/natural*. See also Appendix A for the full questionnaires.

Overall, anthropomorphism has to be interpreted with caution in HRI, in spite of being broadly researched. Due to a lack of a widely understood theoretical basis, the term has been used cover wildly different operationalisations, some of which focus on appearance and form, while others go beyond the visualisation and attempt to manipulate mind attribution. Moreover, possibly because the term is interpreted so broadly, far from all experimental research in HRI that experimentally manipulates anthropomorphism in one way or another pre-tests their manipulation or includes a manipulation check. In general, the field of HRI needs a more stringent and systematic empirical approach to anthropomorphism. Such a framework has been proposed (Epley et al., 2007) and has

even been translated to an HRI context (Eyssel, 2017). Hopefully, this means that it will be just a matter of time before the rest of the field adopts it in their research practices.

7.4 Mind attribution findings

In most of the experiments in this thesis, the attribution of mind to the robot by the participants was measured. A direct comparison of these measures would be invalid due to a number of reasons. One of those reasons is that two different scales were used: one by Kozak et al. (2006) and one by Gray et al. (2007), see Appendix A for the full scales. However, neither study incorporated both, so the extent to which they agree to another cannot be established. In addition, a wide variety of manipulations and robots was used: described versus virtual versus embodied, animated versus still, humanoid versus vehicle-like, etc. Furthermore, mind attribution was successfully manipulated in some experiments, but not in others. As a result, these data sets ought to be considered by themselves rather than combined into one larger data set. However, when similar comparisons within these individual data sets all return the same results, that does suggest that a larger trend exists.

In the sections below, two types of findings will be reported. First, a replication of the two-dimensional model of mind attribution is discussed, for each of the two different scales that were used. Differences between how high robots scored on each of the two scales are reported. Second, overall differences in mind attribution scores between the experiments are discussed; e.g. mind attribution to a virtual versus embodied robot. As mentioned above, no meaningful statistical analyses can be performed on all the accumulated data as there is simply too much variability between the experiments. Still, the descriptives might provide some insights for future research.

7.4.1 Factors of mind attribution

Mind attribution (Gray et al., 2007) consists of two factors, Experience and Agency. Experience is described as possessing "*wants, emotions, and individuality*" (Haslam, Loughnan, et al., 2008, p. 66), while Agency is described as having "*self-control, morality, planning, and thought*" (Haslam, Loughnan, et al., 2008, p. 66). Neither factor in and of itself completely conveys possessing a human mind; only when an agent is perceived as high in both Experience and Agency it is fully humanlike.

Incidentally, the concepts of Experience and Agency overlap substantially with the two factors that encompass dehumanisation theory: Human Nature and Uniquely Human (Bain et al., 2009; Haslam, 2006; Haslam & Loughnan, 2014; M. Y. Li et al., 2014). While Human Nature is "*that which is essential to humanness, the core properties that people share "deep down" despite their superficial variations [...] fundamental, inherent, and natural*"(Haslam, 2006, p. 256), Uniquely Human traits have been defined as "*characteristics [which] define the boundary that separates humans from the related category of animals*" (Haslam, 2006, p. 256).

The two factors have been empirically established for both dehumanisation theory

(Haslam, Loughnan, et al., 2008) and mind attribution (Gray et al., 2007). To further confirm these factors, principal component analysis with two components was conducted on the measures of mind attribution in the experiments described in Experiments I, II, IV and V (Chapters 2, 3, 5, and 6). See Tables 7.3 and 7.4 for the factor loadings on both scales of mind attribution.

A few things stand out when looking at the tables. First of all, in both the Kozak (Table 7.3) and Gray (Table 7.4) scales there is one factor that describes the capacity to experience emotions, feelings, and physical sensations; and one factor that refers to cognition, memory, self-awareness, and the ability to reason about consequences. Thus, the distinction between an Agency factor and an Experience factor can be seen.

However, secondly, factor analysis on the data collected on the scale by Gray (Table 7.4) gave slightly different factors for the different experiments. Not all items on this scale appear to belong firmly to one factor or another. The items 'possessing unique personality traits', 'conveying thoughts and feelings to others', and 'understanding how others are feeling' were classified as Agency in Experiment II, but as components of Experience in Experiment I. In addition, the 'having consciousness' and 'engaging in thought' items in the Kozak scale (Table 7.3) surprisingly did not end up in the Agency factor but rather in Experience. In contrast, in the Gray scale 'having experiences and being aware of things' did get categorised in the Agency factor. Maybe the term 'consciousness' is too broad and was interpreted as simply the capacity to feel, whereas 'having experiences and being aware of things' implies more cognitive processing.

What is maybe most surprising, however, is that these factor structures mostly persisted. Mind attribution was measured to robots that were virtual, videotaped, purely vignette based, humanoid embodied, and non-verbal non-humanoid embodied. In spite of the variability in how the human-robot interaction component was operationalised across the experiments, the two factor structures remained mostly intact.

7.4.2 Mind attribution factors across experiments and different robot embodiments

Due to the two different scales that were used, the differences in robot embodiment, and the difference in how some items were categorised, the experiments cannot be compared with one another. However, it is possible to search for differences between the average score on the two factors for each experiment individually. To do this, for each experiment the average score for each of the two factors of mind attribution was calculated and rescaled so that the score fell in between 0 and 1. According to the literature (Bain et al., 2009; Gray et al., 2007; Haslam, Kashima, Loughnan, Shi, & Suitner, 2008), robots are more readily associated with concepts related to Agency than concepts related to Experience. Thus, it may not be a big surprise that in all experiments, robots were attributed significantly more Agency-related items than Experience-related items. See Table 7.5 for the average score on both factors as well as the difference (Δ) and a significance test.

A second observation stands out. Experiments I and II.A were the only two experiments where participants passively watched (Experiment I) or read about (Experiment

	Experiment IV				Experiment V			
Robot embodiment	*Virtual (pilot)*		*Virtual (main study)*		*Virtual*		*Embodied*	
Item	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*
Engaging in thought	**.62**	.46	**.61**	.51	**.64**	.57	**.61**	.40
Experiencing complex feelings	**.80**	.30	**.81**	.34	**.84**	.31	**.79**	.17
Doing things on purpose	.25	**.70**	.41	**.72**	.22	**.81**	.25	**.72**
Experiencing pain	**.81**	-.01	**.87**	.01	**.86**	.11	**.88**	.08
Having consciousness	**.84**	.30	**.86**	.26	**.89**	.19	**.76**	.31
Having emotion	**.90**	.18	**.83**	.24	**.92**	.23	**.90**	.07
Taking planned action	.04	**.81**	.10	**.84**	.09	**.81**	.04	**.79**
Remembering things	.13	**.81**	.07	**.85**	.09	**.79**	.07	**.75**
Experiencing pleasure	**.86**	.19	**.86**	.21	**.90**	.13	**.88**	.11
Having goals	.46	**.64**	.32	**.69**	.29	**.71**	.53	**.64**

Table 7.3: Factor loadings on the ten items of the mind attribution scale by Kozak et al. (2006) across Experiments IV and V (Chapters 5 and 6).

	Experiment I		Experiment II			
Robot embodiment	*Video*		*Vignette*		*Embodied*	
Item	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*
Feeling hungry	**.86**	.32	**.57**	-.13	**.43**	.15
Feeling afraid or fearful	**.84**	.42	**.73**	.43	**.75**	.19
Experiencing physical or emotional pain	**.85**	.35	**.76**	.29	**.76**	-.03
Experiencing physical or emotional pleasure	**.86**	.39	**.82**	.29	**.77**	.10
Experiencing violent or uncontrolled anger	**.70**	.52	**.74**	.16	**.51**	.31
Longing or hoping for things	**.85**	.42	**.68**	.49	**.61**	.12
Possessing unique personality traits^{x}	**.77**	.48	.48	**.68**	.24	**.72**
Having experiences and being aware of things	.47	**.77**	.35	**.77**	.12	**.78**
Experiencing pride	**.83**	.40	**.73**	.47	**.57**	.06
Experiencing embarrassment	**.88**	.38	**.87**	.20	**.59**	.23
Experiencing joy	**.78**	.47	**.78**	.43	**.83**	.09
Exercising self-restraint over desires, emotions and impulses	.48	**.66**	.32	**.67**	.27	**.36**
Telling right from wrong and doing the right thing	.52	**.75**	.23	**.78**	.37	**.50**
Remembering things	.18	**.91**	-.12	**.61**	-.19	**.62**
Understanding how others are feeling^{x}	**.75**	. 56	.37	**.68**	.17	**.39**
Making plans and working towards goals	.44	**.75**	.13	**.77**	-.08	**.64**
Conveying thoughts and feelings to others^{x}	**.68**	.56	.40	**.61**	.35	**.55**
Thinking	.48	**.78**	.23	**.78**	.33	**.60**

Table 7.4: Factor loadings on the 18 items of the mind attribution scale by Gray et al. (2007) across Experiments I and II (Chapter 2 and 3). Items in italics and marked with ^{x} were not unanimously put in one factor throughout the studies. Factor 1 corresponds with the Experience factor, Factor 2 with Agency.

	Factor 1	*Factor 2*	Δ	t-test	sign.
Experiment I (video)[a]	.18	.42	*.24*	$t(67) = -8.77$	**
Experiment II.A (vignette)[a]	.18	.44	*.26*	$t(192) = -18.22$	**
Experiment II.B (embodied)[a]	.54	.62	*.07*	$t(66) = -2.98$	*
Experiment IV.A (virtual)[b]	.42	.66	*.25*	$t(184) = -14.71$	**
Experiment IV.B (virtual)[b]	.50	.67	*.17*	$t(112) = -8.51$	**
Experiment V (virtual)[b]	.31	.58	*.27*	$t(142) = -12.41$	**
Experiment V (embodied)[b]	.43	.61	*.18*	$t(81) = -6.76$	**

Table 7.5: Differences between Factor 1 (Experience) and Factor 2 (Agency) scores in the different experiments. All scores have been rescaled to lie between 0 and 1. * denotes significance at the $p < .05$ level, ** denotes significance at the $p < .001$ level. Experiments marked with [a] were measured on the scale provided by Gray et al. (2007), experiments marked with [b] on the scale provided by Kozak et al. (2006).

II.A) an interaction between another human and a robot. In all other experiments, participants interacted with a robot (virtual or embodied) themselves. Incidentally, the robots in those two "passive" experiments got the lowest scores on both the Agency and Experience factor. While running a statistical test on this difference would be inappropriate for the reasons outlined above, it might be worthwhile to further investigate whether interaction with social robots enhances the attribution of Experience traits in future research.

Finally, a very tentative observation can be made with regards to the gap size between attributes Experience and Agency. This gap appears to be smaller for embodied robots (Δs = .07 and .18) than for virtual robots (Δs = .17, .24, .25, .26, .27), with one exception (Experiment IV.B). However, caution is warranted when interpreting this difference, due to the variety in scales used, the small number of studies, and the one experiment that breaks the pattern. At most, it should be taken as a potential

7.5 Generalisations and implications

The studies reported in this thesis adopted a variety of robot, interaction, and embodiment types. As discussed above in section 7.4.2 this most likely had consequences for mind attribution and behaviour (as also found in Chapter 6). With that in mind, (how) can the findings in this thesis be generalised to HRI in general?

In a way this question reflects one of the major challenges in the field of HRI. There is a lot of variability between robots, both in terms of hardware and in software. Of course, this can be taken as an argument for only allowing comparisons within specific robot types, so that the "noise" of using different agents will be reduced. However, this solution will not give any meaningful results. Robotics has been developing at an incredible speed and with no signs of slowing down any time soon, so to focus on one specific type of robot might yield less variable results, but those inevitably will become outdated in a few years as robot design evolves and improves. Therefore, in my opinion the only sensible approach is twofold: to base oneself on theoretical frameworks (as already argued above

in section 7.3) and to conduct experiments with as wide a range of agents possible so one can look for the constants in the chaos that ensues. This was done to some extent in this thesis already, which means that the consistent findings on the relation between mind attribution and perceived acceptability of abuse (see sections 7.1.1 and 7.1.2) can most likely be generalised to other robots as well. For the more ambiguous findings, there clearly is a need for further research that will identify which (other) factors are critical in robot bullying behaviour. This need for extensive replication can be seen as a limitation, but then again it is at the base of falsification-based scientific research: one rejection of a null hypothesis in and of itself is not particularly meaningful. Only after multiple replications we can say that the null hypothesis seems (increasingly) unlikely.

In the meantime, a few tentative implications can still be found for practice. Social behaviour may enhance engagement but will not prevent abuse. If anything, reminding the aggressor that the robot is an insentient machine and having the robot refrain from giving off any (social) responses may be the best approach until more is known about the psychological mechanisms that motivate abusive behaviour in the first place. In a way, this is similar to the advice often given to bullying victims "just don't show you care and the bully will leave you alone" (Sokol et al., 2016).

Of course, situations in which this response may not be an option are abundant. For example, delivery robots may not have the time to shut down and patiently wait for a bully to lose interest; especially when amongst other traffic. Even if this disengagement strategy is possible, it will decrease the efficiency of the robot. Previous studies have already opted for alternative discouragement strategies that were found through trial and error (Brscić et al., 2015). Eventually however, we should strive for abuse prevention rather than developing de-escalating responses.

7.6 Future research

Multiple suggestions for future research have come up throughout Experiments I through V. In this section, the most pressing suggestions and directions will be discussed.

As indicated in section 7.3, future research should start to define what aspect of anthropomorphism specifically will be under investigation, if they choose to use the blanket term at all. Moreover, the field will have to develop a habit of including manipulation checks. For these, the researchers will have to clearly state their choice between more psychological measures of anthropomorphism (such as mind attribution), which will take into account individual differences in anthropomorphising tendencies and situational factors; and more objective assessments (such as the revised Godspeed questionnaire) which do not. These choices will impact the conclusions that can (and cannot) be drawn from the experiment; this should be reflected in the conclusion sections.

In spite of the best efforts, this thesis could not give a definite answer to the question how mind perception influences robot bullying behaviour. As discussed in section 7.1.3, this could be because bullying is one of the forms of aggression where an understanding of the victim's mind is essential. Alternatively, it could have been due to an imprecise aggression measurement. A third option is that our repeated failure to manipulate mind

attribution undermined the study design. Even if the results discussed in this thesis had been unanimous, drawing inferences on the nature of the relationship between bullying behaviour and mind attribution would not have been possible. However, the reported experiments do imply that mind attribution is relevant to robot bullying, thus suggesting a promising direction for future research on robot abuse discouraging strategies.

As discussed in section 7.2 it is complicated to conceive an experimental measure of robot bullying. Future research should further develop meaningful measurements and explore alternative motivations of robot bullying, focusing for example on power dynamics and perceived threat or stereotypes held against robots.

An interesting venue of research may be the difference in automatic (implicit, reflexive) and reflective (explicit, cognitive) mind attribution. Especially for robots, this is a fascinating conundrum, as previous research has strongly suggested that people automatically perceive robots as social agents, yet rationally know that they are not. It would be fascinating to see how these two processing mechanisms interact, and what their effect is on emotional and behavioural responses to a robot.

Finally, an obvious lacuna in this thesis is the lack of field research (perhaps with exception of the Cleverbot analysis, Study III). Experiments II, IV and V operationalised bullying behaviour as a tendency to pick a less polite answer, a condescending review, or reduce the power supply to the robot. These operationalisations were founded on the literature, and they helped gain insights in factors that might be related to robot bullying, most notably mind attribution. However, cross-validation of the findings in a field experiment would provide ecological validity.

7.7 Last words

The field of human-robot interaction is young, and highly interdisciplinary. As a result, researchers and engineers are still coming to terms with the vast range of factors that shape HRI — in terms of technical possibilities, human preferences and responses, and ethical implications. The current thesis aimed to shine some light on the niche topic of robot bullying.

The fact that humans perceive and respond to robots as if they were to some extent sentient and humanlike provides the unique opportunity for psychologists to conduct experiments on human interactions where one side of the interaction is completely controlled. However, it can also provide a mirror that shows how some humans behave to those whom they consider subordinate. The fist effects of this potentially confronting view are already visible, in the discussions surrounding whether basic civilities such as saying "thank you" and "please" are necessary in human-AI interaction, and the responses AI assistants should give to sexual harassment. Humans created a new interaction partner; now we have to decide how we want to behave around it.

References

Adobe Systems Software. (2017). *Adobe After Effects CC for MacOS (14.2.1) [computer software].*

Aldebaran Robotics, S. G. (2014). *Choregraphe for MacOS (2.1.4) [computer software].*

Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological science*, *12*(5), 353–359. doi: 10.1111/1467-9280.00366

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., … Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological bulletin*, *136*(2), 151. doi: 10.1037/a0018251

Ang, R. P., & Goh, D. H. (2010). Cyberbullying among adolescents: The role of affective and cognitive empathy, and gender. *Child Psychiatry & Human Development*, *41*(4), 387–397. doi: 10.1007/s10578-010-0176-3

Apple Inc. (1995-2016). *TextEdit (Version 1.12 (329)) [computer software], voice "Princess".*

Ardissono, L., Boella, G., & Lesmo, L. (2000). A plan-based agent architecture for interpreting natural language dialogue. *International Journal of Human-Computer Studies*, *52*(4), 583–635. doi: 10.1006/ijhc.1999.0347

Bain, P., Park, J., Kwok, C., & Haslam, N. (2009). Attributing human uniqueness and human nature to cultural groups: Distinct forms of subtle dehumanization. *Group Processes & Intergroup Relations*, *12*(6), 789-805. doi: 10.1177/1368430209340415

Bartneck, C. (2013). Robots in the theatre and the media. In *Design & semantics of form & movement (desform2013)* (p. 64-70). Philips. Retrieved from `http://www.bartneck.de/publications/2013/robotsTheatreMedia/bartneckDesForm2013.pdf` doi: 10.13140/RG.2.2.28798.79682

Bartneck, C., Duenser, A., Moltchanova, E., & Zawieska, K. (2015). Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. *PloS one*, *10*(4), e0121595. doi: 10.1371/journal.pone.0121595

Bartneck, C., & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies*, *9*(3), 415-433. doi: 10.1075/is.9.3.04bar

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived

safety of robots. *International journal of social robotics*, *1*(1), 71–81. doi: 10.1007/s12369-008-0001-3

Bartneck, C., Reichenbach, J., & Carpenter, J. (2008). The carrot and the stick - the role of praise and punishment in human-robot interaction. *Interaction Studies - Social Behaviour and Communication in Biological and Artificial Systems*, *9*(2), 179-203. doi: 10.1075/is.9.2.03bar

Bartneck, C., Rosalia, C., Menges, R., & Deckers, I. (2005). Robot abuse – a limitation of the media equation. In *Proceedings of the interact 2005 workshop on agent abuse.* Rome, Italy: Designed Intelligence. doi: 10.17605/OSF.IO/4FXQ6

Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007). Daisy, daisy, give me your answer do!: switching off a robot. In *Proceedings of the 2nd ACM/IEEE international conference on human-robot interaction (HRI)* (p. 217-222). Arlington, USA: ACM/IEEE. doi: 10.1145/1228716.1228746

Bartneck, C., Verbunt, M., Mubin, O., & Mahmud, A. A. (2007). To kill a mockingbird robot. In *Proceedings of the 2nd ACM/IEEE international conference on human-robot interaction (HRI)* (p. 81-87). 1179031: ACM Press. doi: 10.1145/1228716.1228728

Bartneck, C., Yogeeswaran, K., Ser, Q. M., Woodward, G., Sparrow, R., Wang, S., & Eyssel, F. (2018). Robots and racism. In *Proceedings of the 13th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 196–204). doi: 10.1145/3171221.3171260

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology*, *6*(1), 1–62. doi: 10.1016/S0065-2601(08)60024-6

Bond, L., Wolfe, S., Tollit, M., Butler, H., & Patton, G. (2007). A comparison of the gatehouse bullying scale and the peer relations questionnaire for students in secondary school. *Journal of School Health*, *77*(2), 75–79. doi: 10.1111/j.1746-1561.2007.00170.x

Boston Dynamics. (2018, February 20). *Testing robustness.* `https://youtu.be/aFuA50H9uek`.

Bosworth, K., Espelage, D. L., & Simon, T. R. (1999). Factors associated with bullying behavior in middle school students. *The journal of early adolescence*, *19*(3), 341–362. doi: 10.1177/0272431699019003003

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Brahnam, S. (2005). Strategies for handling customer abuse of ecas. *Abuse: The darker side of Human-Computer Interaction*, 62–67.

Brahnam, S., & De Angeli, A. (2008). Special issue on the abuse and misuse of social agents. *Interacting with Computers*, *20*, 287–291. doi: 10.1016/j.intcom.2008.02.001

Brahnam, S., & De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers*, *24*(3), 139–153.

Briggs, G., & Scheutz, M. (2014). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social*

Robotics, *6*(3), 343-355. doi: 10.1007/978-3-642-34103-8_24

Brscić, D., Kidokoro, H., Suehiro, Y., & Kanda, T. (2015). Escaping from children's abuse of social robots. In *Proceedings of the 10th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 59–66). Portland, USA: ACM/IEEE.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi: 10.1177/1745691610393980

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. New York: Springer Science & Business Media. doi: 10.1198/tech.2003.s147

Calverley, D. J. (2006). Android science and animal rights, does an analogy exist? *Connection Science*, *18*(4), 403–417. doi: 10.1080/09540090600879711

Casper, D. M., Meter, D. J., & Card, N. A. (2015). Addressing measurement issues related to bullying involvement. *School Psychology Review*, *44*(4), 353–371. doi: 10.17105/spr-15-0036.1

Castano, E., & Kofta, M. (2009). Dehumanization: Humanity and its denial. *Group Processes & Intergroup Relations*, *12*(6), 695-697. doi: 10.1177/1368430209350265

Castano, E., Kofta, M., Čehajić, S., Brown, R., & González, R. (2009). What do I care? Perceived ingroup responsibility and dehumanization as predictors of empathy felt for the victim group. *Group Processes & Intergroup Relations*, *12*(6), 715-729. doi: 10.1177/1368430209347727

Chin, H., & Yi, M. Y. (2019). Should an agent be ignoring it?: A study of verbal abuse types and conversational agents' response styles. In *Extended abstracts of the acm conference on human factors in computing systems (CHI)* (p. 1-6). doi: 10.1145/3290607.3312826

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155. doi: 10.1037/0033-2909.112.1.155

Cooper, A. (2019, June). *How robots change the world; what automation really means for jobs and productivity* (Tech. Rep.). Oxford: Oxford Economics.

Cowie, H., & Berdondini, L. (2002). The expression of emotion in response to bullying. *Emotional and Behavioural Difficulties*, *7*(4), 207–214.

Crick, N. R., & Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child development*, *66*(3), 710–722. doi: 10.1111/j.1467-8624.1995.tb00900.x

Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, *24*(6), 588-594. doi: 10.1177/0146167298246003

Cronbach, L. J. (1951, Sep 01). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. doi: 10.1007/BF02310555

Curry, A. C., & Rieser, V. (2018). #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the second acl workshop on ethics in natural*

language processing (pp. 7–14). doi: 10.18653/v1/W18-0802

Dahl, J., Vescio, T., & Weaver, K. (2015). How threats to masculinity sequentially cause public discomfort, anger, and ideological dominance over women. *Social Psychology*. doi: 10.1027/1864-9335/a000248

Darling, K. (2012). Extending legal rights to social robots. In *We Robot Conference, University of Miami* (p. 1-24). Miami, USA: University of Miami. doi: 10.2139/ssrn.2044797

Darling, K. (2015). 'Who's Johnny?'Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. *Robot Ethics*, *2*. doi: 10.2139/ssrn.2588669

Darling, K., Nandy, P., & Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. In *Proceedings of the 24th IEEE International symposium on robot and human interactive communication (RO-MAN)* (pp. 770–775). doi: 10.1109/ROMAN.2015.7333675

David, B., Grace, D., & Ryan, M. K. (2004). The gender wars: A self-categorisation theory perspective on the development of gender identity. In *The development of the social self* (pp. 149–172). Psychology Press.

De Angeli, A. (2006). On verbal abuse towards chatterbots. In *Proceedings of CHI workshop on misuse and abuse of interactive technologies.* doi: 10.1016/j.intcom.2008.02.004

De Angeli, A. (2009). Ethical implications of verbal disinhibition with conversational agents. *PsychNology Journal*, *7*(1), 49-57.

De Angeli, A., & Brahnam, S. (2006). Sex stereotypes and conversational agents. In *Gender and interaction: real and virtual women in a male world* (p. 1-4).

De Angeli, A., & Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with computers*, *20*(3), 302-310. doi: 10.1016/j.intcom.2008.02.004

De Angeli, A., Brahnam, S., Wallis, P., & Dix, A. (2006). Misuse and abuse of interactive technologies. In *Extended abstracts on human factors in computing systems (CHI)* (p. 1647-1650). Montreal, Canada: ACM. doi: 10.1145/1125451.1125753

De Angeli, A., & Carpenter, R. (2005). Stupid computer! Abuse and social identities. In *Proceedings interact 2005 workshop abuse: The darker side of human-computer interaction* (pp. 19–25).

De Angeli, A., Johnson, G. I., & Coventry, L. (2001). The unfriendly user: exploring social reactions to chatterbots. In *Proceedings of the international conference on affective human factors design, london* (pp. 467–474).

De Swert, K. (2012). Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha [Computer software manual]. University of Amsterdam, the Netherlands.

Dindia, K., Fitzpatrick, M. A., & Kenny, D. A. (1997). Self-disclosure in spouse and stranger interaction: A social relations analysis. *Human Communication Research*, *23*(3), 388–412. doi: 10.1111/j.1468-2958.1997.tb00402.x

Dunning, D. (1999). A newer look: Motivated social cognition and the schematic representation of social concepts. *Psychological Inquiry*, *10*(1), 1–11. doi: 10.1207/

s15327965pli1001_1

Ensari, N., & Miller, N. (2002). The out-group must not be so bad after all. The effects of disclosure, typicality, and salience on intergroup bias. *Journal of Personality and Social Psychology*, *83*(2), 313. doi: 10.1037/0022-3514.83.2.313

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science*, *19*(2), 114-120. doi: 10.1111/j.1467-9280 .2008.02056.x

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864-886. doi: 10.1037/0033 -295X.114.4.864

Eyssel, F. (2017). An experimental psychological perspective on social robotics. *Robotics and Autonomous Systems*, *87*, 363-371. doi: 10.1016/j.robot.2016.08.029

Eyssel, F., & Hegel, F. (2012). (S)he's got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology*, *42*(9), 2213–2230. doi: 10.1111/j.1559-1816 .2012.00937.x

Eyssel, F., & Kuchenbrandt, D. (2011). Manipulating anthropomorphic inferences about nao: The role of situational and dispositional aspects of effectance motivation. In *2011 ro-man* (pp. 467–472). doi: 10.1109/ROMAN.2011.6005233

Eyssel, F., Kuchenbrandt, D., & Bobinger, S. (2011). Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of the 6th international conference on human-robot interaction* (pp. 61–68). doi: 10.1145/1957656.1957673

Eyssel, F., Kuchenbrandt, D., Bobinger, S., de Ruiter, L., & Hegel, F. (2012). 'If you sound like me, you must be more human': on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the 7th annual ACM/IEEE international conference on human-robot interaction (HRI)* (p. 125-126). Boston, USA: ACM/IEEE. doi: 10.1145/2157689.2157717

Eyssel, F., Kuchenbrandt, D., Hegel, F., & de Ruiter, L. (2012). Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In *2012 ieee ro-man: The 21st ieee international symposium on robot and human interactive communication* (pp. 851–857). doi: 10.1109/ROMAN.2012.6343858

Eyssel, F., & Pfundmair, M. (2015). Predictors of psychological anthropomorphization, mind perception, and the fulfillment of social needs: A case study with a zoomorphic robot. In *Proceedings of the 24th IEEE international symposium on robot and human interactive communication (RO-MAN)* (p. 827-832). Kobe, Japan: IEEE. doi: 10 .1109/ROMAN.2015.7333647

Eyssel, F., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots)—on the effects of loneliness on psychological anthropomorphism. In *2013 8th acm/ieee international conference on human-robot interaction (hri)* (pp. 121–122). doi: 10 .1109/HRI.2013.6483531

Facchin, F., Barbara, G., & Cigoli, V. (2017). Sex robots: the irreplaceable value of

humanity. *BMJ*, *358*, j3790.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175–191. doi: 10.3758/BF03193146

Ferguson, C. J. (2007). The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric quarterly*, *78*(4), 309–316. doi: 10.1007/s11126-007-9056-9

Fessler, L. (2017a, December). *Apple and amazon are under fire for siri and alexa's responses to sexual harassment.* `https://qz.com/work/1151282/siri-and-alexa-are-under-fire-for-their-replies-to-sexual-harassment/`.

Fessler, L. (2017b, February). *We tested bots like siri and alexa to see who would stand up to sexual harassment.* `https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/`.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Sage publications.

Fisher, J. A. (1995). The myth of anthropomorphism. In *Readings in animal cognition* (pp. 3–16). MIT Press.

Fox, C. L., & Boulton, M. J. (2005). The social skills problems of victims of bullying: Self, peer and teacher perceptions. *British Journal of Educational Psychology*, *75*(2), 313–328.

Franklin Jr, R. G., Nelson, A. J., Baker, M., Beeney, J. E., Vescio, T. K., Lenz-Watson, A., & Adams Jr, R. B. (2013). Neural responses to perceiving suffering in humans and animals. *Social neuroscience*, *8*(3), 217–227. doi: 10.1080/17470919.2013.763852

Frith, C. D., & Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron*, *60*(3), 503–510. doi: 10.1016/j.neuron.2008.10.032

Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *Journal of Personality and Social Psychology*, *85*(3), 453-466. doi: 10.1037/0022-3514.85.3.453

Galinsky, A. D., Magee, J. C., Inesi, M. E., & Gruenfeld, D. H. (2006). Power and perspectives not taken. *Psychological Science*, *17*(12), 1068-1074. doi: 10.1111/j.1467-9280.2006.01824.x

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage*, *35*(4), 1674-1684. doi: 10.1016/j.neuroimage.2007.02.003

Goodwin, S. A., Gubin, A., Fiske, S. T., & Yzerbyt, V. Y. (2000). Power can bias impression processes. stereotyping subordinates by default and by design. *Group Processes & Intergroup Relations*, *3*(3), 227–256.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. doi: 10.1126/science.1134475

Gruenfeld, D. H., Inesi, M. E., Magee, J. C., & Galinsky, A. D. (2008). Power and the objectification of social targets. *Journal of personality and social psychology*, *95*(1), 111. doi: 10.1037/0022-3514.95.1.111

Gullone, E., & Robertson, N. (2008). The relationship between bullying and animal abuse behaviors in adolescents: The importance of witnessing animal abuse. *Journal of Applied Developmental Psychology*, *29*(5), 371–379. doi: 10.1016/j.appdev.2008.06.004

Gwinn, J. D., Judd, C. M., & Park, B. (2013). Less power= less human? effects of power differentials on dehumanization. *Journal of Experimental Social Psychology*, *49*(3), 464-470. doi: 10.1016/j.jesp.2013.01.005

Ham, J., & Midden, C. (2009). A robot that says bad! Using negative and positive social feedback from a robotic agent to save energy. In *Proceedings of the 4th ACM/IEEE international conference on human robot interaction (HRI)* (pp. 265–266). doi: 10.1145/1514095.1514168

Hamburger, M. E., Basile, K. C., & Vivolo, A. M. (2011). *Measuring bullying victimization, perpetration, and bystander experiences; a compendium of assessment tools.* Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252-264. doi: 10.1207/s15327957pspr1003_4

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social Cognition*, *26*(2), 248–258. doi: 10.1521/soco.2008.26.2.248

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*, 399-423. doi: 10.1146/annurev-psych-010213-115045

Haslam, N., Loughnan, S., Kashima, Y., & Bain, P. (2008). Attributing and denying humanness to others. *European review of Social Psychology*, *19*(1), 55-85. doi: 10.1080/10463280801981645

Hein, G., & Singer, T. (2008). I feel how you feel but not always: the empathic brain and its modulation. *Current opinion in neurobiology*, *18*(2), 153–158. doi: 10.1016/j.conb.2008.07.012

Hern, A. (2010, September). *Apple made siri deflect questions on feminism, leaked papers reveal.* `https://www.theguardian.com/technology/2019/sep/06/apple-rewrote-siri-to-deflect-questions-about-feminism`. ([Online; recovered 23 June 2020])

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, *49*, 245–250. doi: 10.1016/j.chb.2015.02.026

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, *68*(4), 712–733. doi: 10.1093/joc/jqy026

Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. *Computers in Human Behavior*, *26*(6), 1508-1518. doi: 10.1016/j.chb.2010.05.015

Horstmann, A. C., Bock, N., Linhuber, E., Szczuka, J. M., Straßmann, C., & Krämer, N. C. (2018). Do a robot's social skills and its objection discourage interactants from switching the robot off? *PloS one*, *13*(7), e0201581. doi: 10.1371/journal.pone .0201581

Hutchinson, M. K., & Holtman, M. C. (2005). Analysis of count data using Poisson regression. *Research in Nursing & Health*, *28*(5), 408–418.

Ireland, J. L., & Monaghan, R. (2006). Behaviours indicative of bullying among young and juvenile male offenders: A study of perpetrator and victim characteristics. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *32*(2), 172–180.

Jolliffe, D., & Farrington, D. P. (2011). Is low empathy related to bullying after controlling for individual and social background variables? *Journal of Adolescence*, *34*(1), 59–71. doi: 10.1016/j.appdev.2008.06.004

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., … Shen, S. (2012). "robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, *48*(2), 303. doi: 10.1037/a0027033

Kanda, T., Sato, R., Saiwaki, N., & Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on robotics*, *23*(5), 962–971.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*. doi: 10.3389/ fpsyg.2015.00390

Keijsers, M., & Bartneck, C. (2018). Mindless robots get bullied. In *Proceedings of the 13th ACM/IEEE international conference on human-robot interaction (HRI)* (p. 205-214). New York, USA. doi: 10.1145/3171221.3171266

Keijsers, M., Bartneck, C., & Kazmi, H. S. (2019). Cloud-based sentiment analysis for interactive agents. In *Proceedings of the 7th international conference on human-agent interaction (HAI)* (pp. 43–50). doi: 10.1145/3349537.3351883

Keijsers, M., Kazmi, H., Eyssel, F., & Bartneck, C. (2019). Teaching robots a lesson: Determinants of robot punishment. *International Journal of Social Robotics*, 1–14. doi: 10.1007/s12369-019-00608-w

Kim, S., & McGill, A. L. (2011). Gaming with Mr. Slot or gaming the slot machine? Power, anthropomorphism, and risk perception. *Journal of Consumer Research*, *38*(1), 94-107. doi: 10.1086/658148

Kozak, M. N., Marsh, A. A., & Wegner, D. M. (2006). What do i think you're doing? action identification and mind attribution. *Journal of Personality and Social Psychology*, *90*(4), 543-555. doi: 10.1037/0022-3514.90.4.543

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? interaction and perspective taking with robots investigated via fmri. *PloS One*, *3*(7), e2597. doi: 10.1371/journal.pone.0002597

Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *Journal of Personality and Social Psychology*, *109*(5), 901. doi: 10.1037/pspp0000048

Ku, H., Choi, J. J., Lee, S., Jang, S., & Do, W. (2018). Designing shelly, a robot capable of assessing and restraining children's robot abusing behaviors. In *Companion of the 13th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 161–162). doi: 10.1145/3173386.3176973

Kuchenbrandt, D., Riether, N., & Eyssel, F. (2014). Does anthropomorphism reduce stress in hri? In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction* (pp. 218–219). doi: 10.1145/2559636.2563710

Lammers, J., & Stapel, D. A. (2011). Power increases dehumanization. *Group Processes & Intergroup Relations*, *14*(1), 113-126. doi: 10.1177/1368430210370042

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*(2), 434-443. doi: 10.1016/j.chb.2011.10.014

Lasica, J. D. (2014). *Knightscope K5 at the Launch Festival, held Feb. 24-26, 2014 at San Francisco's Design Concourse.* `https://commons.wikimedia.org/wiki/File:Knightscope_K5_(12809731473).jpg`. ([Online; recovered 25 June 2019])

Lee, K. M. (2004). Presence, explicated. *Communication Theory*, *14*(1), 27-50. doi: 10.1111/j.1468-2885.2004.tb00302.x

Leidner, B., Castano, E., & Ginges, J. (2013). Dehumanization, retributive and restorative justice, and aggressive versus diplomatic intergroup conflict resolution strategies. *Personality and Social Psychology Bulletin*, *39*(2), 181-192. doi: 10.1177/0146167212472208

Lesher, J. H. (2001). *Xenophanes of colophon: fragments: a text and translation with a commentary* (Vol. 4). Toronto: University of Toronto Press.

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23-37. doi: 10.1016/j.ijhcs.2015.01.001

Li, M. Y., Leidner, B., & Castano, E. (2014). Toward a comprehensive taxonomy of dehumanization: integrating two senses of humanness, mind perception theory, and stereotype content model. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, *21*(3), 285-300. doi: 10.4473/TPM21.3.4

Liepelt, R., & Brass, M. (2010). Top-down modulation of motor priming by belief about animacy. *Experimental psychology*. doi: 10.1027/1618-3169/a000028

Liepelt, R., Cramon, D., & Brass, M. (2008). What is matched in direct matching? Intention attribution modulates motor priming. *Journal of Experimental Psychology: human perception and performance*, *34*(3), 578. doi: 10.1037/0096-1523.34.3.578

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A monte carlo approach. *Journal of Applied Psychology*, *60*(1), 10. doi: 10.1037/h0076268

Locke, K. D. (2009). Aggression, narcissism, self-esteem, and the attribution of desirable

and humanizing traits to self versus others. *Journal of Research in Personality*, *43*(1), 99-102. doi: 10.1016/j.jrp.2008.10.003

Lortie, C. L., & Guitton, M. J. (2011). Judgment of the humanness of an interlocutor is in the eye of the beholder. *PLoS One*, *6*(9), e25085. doi: 10.1371/journal.pone.0025085

Loughnan, S., & Haslam, N. (2007). Animals and androids: Implicit associations between social categories and nonhumans. *Psychological Science*, *18*(2), 116-121. doi: 10 .1111/j.1467-9280.2007.01858.x

Loughnan, S., Haslam, N., Murnane, T., Vaes, J., Reynolds, C., & Suitner, C. (2010). Objectification leads to depersonalization: The denial of mind and moral concern to objectified others. *European Journal of Social Psychology*, *40*(5), 709-717. doi: 10.1002/ejsp.755

Lowry, P. B., Zhang, J., Wang, C., & Siponen, M. (2016). Why do adults engage in cyberbullying on social media? an integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research*, *27*(4), 962-986. doi: 10.1287/isre.2016.0671

Lucas, H., Poston, J., Yocum, N., Carlson, Z., & Feil-Seifer, D. (2016). Too big to be mistreated? Examining the role of robot size on perceptions of mistreatment. In *Proceedings of the 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 1071–1076). doi: 10.1109/ROMAN.2016.7745241

Luczak, H., Roetting, M., & Schmidt, L. (2003). Let's talk: anthropomorphization as means to cope with stress of interacting with technical devices. *Ergonomics*, *46*(13-14), 1361-1374. doi: 10.1080/00140130310001610883

MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition*, *146*, 190-205. doi: 10.1016/j.cognition.2015.09.019

Malle, B. F., Bello, P., & Scheutz, M. (2019). Requirements for an artificial agent with norm competence. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 21–27). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/3306618.3314252` doi: 10.1145/3306618.3314252

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the 10th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 117–124). ACM. doi: 10.1145/2696454.2696458

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? the impact of action and appearance on people's evaluations of a moral robot. In *Proceedings of the 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 125–132). doi: 10.1109/HRI.2016.7451743

Mauldin, M. L. (1994). Chatterbots, tinymuds, and the Turing test: Entering the Loebner prize competition. In *AAAI* (Vol. 94, pp. 16–21).

Mishna, F., Schwan, K. J., Lefebvre, R., Bhole, P., & Johnston, D. (2014). Students in distress: Unanticipated findings in a cyber bullying study. *Children and Youth Services Review*, *44*, 341–348.

Modecki, K. L., Minchin, J., Harbaugh, A. G., Guerra, N. G., & Runions, K. C. (2014). Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. *Journal of Adolescent Health*, *55*(5), 602-611. doi: 10.1016/j.jadohealth.2014.06.007

Moore, S. (2018, Feb). *Gartner says 25 percent of customer service operations will use virtual customer assistants by 2020.* Press Release. Retrieved from `https://www.gartner.com/en/newsroom/press-releases/2018-02-19-gartner-says-25-percent-of-customer-service-operations-will-use-virtual-customer-assistants-by-2020`

Mori, M., et al. (1970). The uncanny valley. *Energy*, *7*(4), 33–35. doi: 10.1109/MRA.2012.2192811

Moshkina, L., Trickett, S., & Trafton, J. G. (2014). Social engagement in public places: a tale of one robot. In *Proceedings of the 9th ACM/IEEE international conference on human-robot interaction (HRI)* (p. 382-389). Bielefeld, Germany: ACM/IEEE. doi: 10.1145/2559636.2559678

Müller, B. C., van Baaren, R. B., van Someren, D. H., & Dijksterhuis, A. (2014). A present for pinocchio: On when non-biological agents become real. *Social Cognition*, *32*(4), 381–396. doi: 10.1521/soco.2014.32.4.381

Mutlu, B., & Forlizzi, J. (2008). Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In *Proceedings of the 3rd ACM/IEEE international conference on human robot interaction (HRI)* (pp. 287–294). doi: 10.1145/1349822.1349860

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the sigchi conference on human factors in computing systems* (p. 72-78). Boston, USA: ACM. doi: 10.1145/191666.191703

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. doi: 10.1002/ejsp.2420150303

Neyer, F. J., Felber, J., & Gebhardt, C. (2012). Entwicklung und validierung einer kurzskala zur erfassung von technikbereitschaft. *Diagnostica*. doi: 10.1026/0012-1924/a000067

Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., & Yamada, S. (2016). Why do children abuse robots? *Interaction Studies*, *17*(3), 347–369. doi: 10.1075/is.17.3.02nom

Oberman, L. M., McCleery, J. P., Ramachandran, V. S., & Pineda, J. A. (2007). Eeg evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*, *70*(13-15), 2194–2203. doi: 10.1016/j.neucom.2006.02.024

Paetzel, M., Peters, C., Nyström, I., & Castellano, G. (2016). Congruency matters – how ambiguous gender cues increase a robot's uncanniness. In *International conference on social robotics* (pp. 402–412). doi: 10.1007/978-3-319-47437-3_39

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC 2015 [computer software]. Pennebaker Conglomerates.* Retrieved from `http://liwc.wpengine.com`

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (Tech. Rep.). The University of Texas at Austin. Retrieved from `http://hdl.handle.net/2152/31333`

Postigo, S., González, R., Montoya, I., & Ordoñez, A. (2013). Theoretical proposals in bullying research: a review. *Anales de psicología*, *29*(2).

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. In *Proceedings of the 2nd ACM/IEEE international conference on human-robot interaction (HRI)* (p. 145-152). Arlington, USA: ACM/IEEE. doi: 10.1145/1228716.1228736

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526. doi: 10.1017/S0140525X00076512

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences*, *25*(1), 1–20. doi: 10.1017/S0140525X02000018

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of general psychology*, *14*(2), 154–166. doi: 10.1037/a0019440

Reeves, B., & Nass, C. (1996). *The Media Equation*. Cambridge: CSLI Publications and Cambridge University Press.

Rehm, M., & Krogsager, A. (2013). Negative affect in human robot interaction – impoliteness in unexpected encounters with robots. In *Proceedings of the 22nd IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 45–50). doi: 10.1109/ROMAN.2013.6628529

Reich, N., & Eyssel, F. (2013). Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables. *Paladyn, Journal of Behavioral Robotics*, *4*(2), 123–130. doi: 10.2478/pjbr-2013-0014

Reichenbach, J., Bartneck, C., & Carpenter, J. (2006). Well done, robot! the importance of praise and presence in human-robot collaboration. In *Proceedings of the 15th IEEE international symposium on robot and human interactive communication (RO-MAN)* (p. 86-90). Hatfield, UK: IEEE. doi: 10.1109/ROMAN.2006.314399

Reich-Stiebert, N., & Eyssel, F. (2017). (ir) relevance of gender? on the influence of gender stereotypes on learning with a robot. In *2017 12th acm/ieee international conference on human-robot interaction (hri* (pp. 166–176). doi: 10.1145/2909824.3020242

Richardson, K. (2016). The asymmetrical'relationship': parallels between prostitution and the development of sex robots. *ACM SIGCAS Computers and Society*, *45*(3), 290-293. doi: 10.1145/2874239.2874281

Richter, S. J., & Richter, C. (2002). A method for determining equivalence in industrial applications. *Quality Engineering*, *14*(3), 375–380. doi: 10.1081/QEN-120001876

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on human-robot interaction (HRI)* (p. 245-246). San Diego, USA: ACM/IEEE. doi: 10.1145/1514095.1514158

Ritter, D., & Eslea, M. (2005). Hot sauce, toy guns, and graffiti: A critical account of

current laboratory aggression paradigms. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *31*(5), 407–419. doi: 10.1002/ab.20066

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, *5*(1), 17-34. doi: 10.1007/s12369-012-0173-8

Rosenthal-Von Der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., … Krämer, N. C. (2014). Investigations on empathy towards humans and robots using fmri. *Computers in Human Behavior*, *33*, 201–212. doi: 10.1016/j.chb.2014.01.004

Rudman, L. A., & Mescher, K. (2012). Of animals and objects: Men's implicit dehumanization of women and likelihood of sexual aggression. *Personality and Social Psychology Bulletin*, *38*(6), 734-746. doi: 10.1177/0146167212436401

Ruijten, P. A., Bouten, D. H., Rouschop, D. C., Ham, J., & Midden, C. J. (2014). Introducing a rasch-type anthropomorphism scale. In *Proceedings of the 9th ACM/IEEE international conference on human-robot interaction (HRI)* (p. 280-281). Bielefeld, Germany: ACM/IEEE. doi: 10.1145/2559636.2559825

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, *5*(3), 313-323. doi: 10.1007/s12369-013-0196-9

Salvini, P., Ciaravella, G., Yu, W., Ferri, G., Manzi, A., Mazzolai, B., … Dario, P. (2010). How safe are service robots in urban environments? bullying a robot. In *Proceedings of the 19th IEEE international symposium on robot and human interactive communication (RO-MAN)* (p. 1-7). Viareggio, Italy: IEEE. doi: 10.1109/ROMAN.2010.5654677

Sandoval, E. B., Brandstetter, J., & Bartneck, C. (2016). Can a robot bribe a human? the measurement of the dark side of reciprocity in human robot interaction. In *Proceedings of the 11th ACM/IEEE international conference on human-robot interaction (HRI)* (p. 117 - 124). Christchurch: IEEE. doi: 10.1109/HRI.2016.7451742

Schulman-Green, D. (2003). Coping mechanisms of physicians who routinely work with dying patients. *OMEGA-Journal of Death and Dying*, *47*(3), 253–264. doi: 10.2190/950H-U076-T5JB-X6HN

Shrauger, J. S. (1975). Responses to evaluation as a function of initial self-perceptions. *Psychological bulletin*, *82*(4), 581. doi: 10.1037/h0076791

Simons, D. J., & Chabris, C. F. (2012). Common (mis)beliefs about memory: A replication and comparison of telephone and mechanical turk survey methods. *PloS one*, *7*(12), e51876. doi: 10.1371/journal.pone.0051876

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., … Sanchez-Vives, M. V. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PloS one*, *1*(1), e39. doi: 10.1371/journal.pone.0000039

Sokol, N., Bussey, K., & Rapee, R. M. (2016). Victims' responses to bullying: The gap between students' evaluations and reported responses. *School Mental Health*, *8*(4),

461–475.

Sparrow, R. (2016). Kicking a robot dog. In *Proceedings of the 11th ACM/IEEE international conference on human-robot interaction (HRI)* (p. 229-229). Christchurch: IEEE. doi: 10.1109/HRI.2016.7451756

Sparrow, R. (2017). Robots, rape, and representation. *International Journal of Social Robotics*, *9*(4), 465-477. doi: 10.1007/s12369-017-0413-z

Stenzel, A., Chinellato, E., Bou, M. A. T., del Pobil, Á. P., Lappe, M., & Liepelt, R. (2012). When humanoid robots become human-like interaction partners: corepresentation of robotic actions. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1073. doi: 10.1037/a0029493

Strait, M., Contreras, V., & Vela, C. D. (2018). Verbal disinhibition towards robots is associated with general antisociality. *arXiv e-prints*.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643. doi: 10.1037/h0054651

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*(3), 321-326. doi: 10.1089/1094931041291295

Sutton, J., Smith, P. K., & Swettenham, J. (1999). Bullying and 'theory of mind': A critique of the 'social skills deficit'view of anti-social behaviour. *Social Development*, *8*(1), 117–127. doi: 10.1111/1467-9507.00083

Tan, X. Z., Vázquez, M., Carter, E. J., Morales, C. G., & Steinfeld, A. (2018). Inducing bystander interventions during robot abuse with social mechanisms. In *Proceedings of the 13th ACM/IEEE international conference on human-robot interaction (HRI)* (p. 169-177). New York, USA: ACM/IEEE. doi: 10.1145/3171221.3171247

Thellman, S., Silvervarg, A., Gulz, A., & Ziemke, T. (2016). Physical vs. virtual agent embodiment and effects on social interaction. In *Proceedings of the 16th international conference on intelligent virtual agents (IVA)* (p. 412-415). Los Angeles, USA: Springer International Publishing. doi: 10.1007/978-3-319-47665-0_44

Urquiza-Haas, E. G., & Kotrschal, K. (2015). The mind behind anthropomorphic thinking: attribution of mental states to other species. *Animal Behaviour*, *109*, 167–176. doi: 10.1016/j.anbehav.2015.08.011

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers*, *20*(3), 292-301. doi: 10.1016/j.intcom.2008.02.007

Vincent, J. (2017, April). *A drunk man was arrested for knocking over Silicon Valley's crime-fighting robot.* `https://www.theverge.com/2017/4/26/15432280/security-robot\-knocked-over-drunk-man-knightscope-k5-mountain-view`. ([Online; recovered 30 August 2018])

Volk, A. A., Veenstra, R., & Espelage, D. L. (2017). So you want to study bullying? recommendations to enhance the validity, transparency, and compatibility of bullying research. *Aggression and Violent Behavior*, *36*, 34–43. doi: 10.1016/j.avb.2017.07.003

Wallis, P. (2005). Robust normative systems: What happens when a normative sys-

tem fails. In *Abuse: the darker side of human-computer interaction, interact 2005* (p. 68-72). Rome, Italy.

Walsh, T. (2018). *2062: The world that AI made*. Black Inc.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? the stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. doi: 10.1177/1745691610369336

Waytz, A., & Epley, N. (2012). Social connection enables dehumanization. *Journal of Experimental Social Psychology*, *48*(1), 70-76. doi: 10.1016/j.jesp.2011.07.012

Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, *19*(1), 58-62. doi: 10.1177/0963721409359302

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410-465. doi: 10.1037/a0020240

Whitby, B. (2008). Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers*, *20*(3), 326-333. doi: 10.1016/j.intcom.2008.02.002

Whiten, A., & Byrne, R. (1991). *Natural theories of mind: Evolution, development and simulation of everyday mindreading*. Basil Blackwell Oxford.

Wiese, E., Mandell, A., Shaw, T., & Smith, M. (2019). Implicit mind perception alters vigilance performance because of cognitive conflict processing. *Journal of experimental psychology: applied*, *25*(1), 25. doi: 10.1037/xap0000186

Wullenkord, R., Fraune, M. R., Eyssel, F., & Šabanović, S. (2016). Getting in touch: How imagined, actual, and physical contact affect evaluations of robots. In *Proceedings of the 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (p. 980-985). New York, USA: IEEE. doi: 10.1109/ROMAN.2016.7745228

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human-Robot Interaction*, *5*(2), 29–47. doi: 10.5898/JHRI.5.2.Yogeeswaran

Young, M. (2016, March 16). *What is a robot?* `https://www.youtube.com/watch?v=S5miA6jXf0E&frags=pl%2Cwn`.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Research Report Series / Department of Statistics and Mathematics*. doi: 10.18637/jss.v011.i10

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, *4*(7).

Złotowski, J., Strasser, E., & Bartneck, C. (2014). Dimensions of anthropomorphism: from humanness to humanlikeness. In G. Sagerer (Ed.), *Proceedings of the 9th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 66–73). New York,

USA: ACM/IEEE.

Złotowski, J., Sumioka, H., Bartneck, C., Nishio, S., & Ishiguro, H. (2017). *Understanding anthropomorphism: Anthropomorphism is not a reverse process of dehumanization* [Unpublished Work].

Złotowski, J., Sumioka, H., Eyssel, F., Nishio, S., Bartneck, C., & Ishiguro, H. (2018). Model of dual anthropomorphism: The relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics*, *10*(5), 701–714. doi: 10.1007/s12369-018-0476-5

Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies*, *100*, 48–54. doi: 10.1016/j.ijhcs.2016.12.008

Appendix A

Questionnaires

Mind attribution

Mind attribution was measured with the 10-item questionnaire by Kozak et al. (2006) in the experiments in Chapter 5 and 6, and with the 18-item scale developed by Gray et al. (2007) in the experiments in the Chapters 2 and 3.

The Mind Attribution Scale by Kozak et al.

Please indicate what extent you agree with the following propositions. I feel like the robot[1] was capable...
of experiencing complex feelings
of feeling pain
of experiencing emotion
of feeling pleasure
of doing things on purpose
of undertaking planned actions
of having goals
of consciousness
of remembering
of engaging in thought

Dimensions of Mind Perception by Gray et al.

Please indicate to what extent this agent would be capable of the following attributes:
Feeling hungry
Feeling afraid of fearful
Conveying thoughts or feelings to others
Having experiences and being aware of things
Experiencing embarrassment

[1] *In original: "This person"*

Understanding how others are feeling
Experiencing joy
Remembering things
Telling right from wrong and trying to do the right thing
Experiencing physical or emotional pain
Having personality traits that make it unique from others
Making plans and working towards a goal
Experiencing physical or emotional pleasure
Experiencing pride
Experiencing violent or uncontrolled anger
Exercising self-restraint over desires, emotions, or impulses
Thinking

Humanlikeness

Humanlikeness was measured in the experiment described in Chapter 5 through the Revised Godspeed Questionnaire by C.-C. Ho and MacDorman (2010).

I thought the robot seemed...	
Artificial	Natural
Synthetic	Real
Living	Inanimate
Human-made	Human-like
Moving naturally	Mechanical
Without definite life span	Mortal

Trait anthropomorphism

Individual differences in anthropomorphism was measured in the experiments described in Chapter 2 and 3. The scale was developed by Waytz, Cacioppo, and Epley (2010).

Please rate the following items. To what extent...
does the average fish have free will?
does the average mountain have free will?
do cows have intentions?
does the ocean have consciousness?
does a cheetah experience emotions?
does the environment experience emotions?
does the average insect have a mind of its own?
does a tree have a mind of its own?
does the wind have intentions?
does the average reptile have consciousness?
In addition, the original scale also included the following items regarding anthropomorphism of technology:

does technology—devices and machines for manufacturing, entertainment, and productive processes (e.g., cars, computers, television sets)—have intentions?

does a television set experience emotions?

does the average robot have consciousness?

does a car have free will?

does the average computer have a mind of its own?

These were removed since the ratings were influenced by the condition participants had been assigned to.

Affinity with technology

The scale was translated from German, with the original being reported in Neyer et al. (2012). It was used in the experiments described in Chapter 3.

How descriptive are the following propositions of you?
I am very curious about new technological developments
I tend to quickly embrace new technology
I am always interested in using the latest technological gadgets
If I had the opportunity, I'd use technology even more often that I currently do
[*reverse coded*] I am often afraid to fail when dealing with modern technology
[*reverse coded*] I find dealing with technological innovations often too demanding
[*reverse coded*] When handling new technology, I am afraid to break it rather than using it the right way
[*reverse coded*] Dealing with new technology is hard for me – I usually simply cannot manage it

Perceived threat from robots

The scale was adopted from Yogeeswaran et al. (2016); Złotowski, Yogeeswaran, and Bartneck (2017). It was used in Experiment V, Chapter 6.

Please indicate your feelings toward the following statements:
The increased use of robots in our everyday life is causing job losses for humans
[*reverse coded*] Robots are not displacing workers from their jobs
In the long run, robots pose a direct threat to human safety and well-being
Advancements in robot technology threaten human employment and opportunities

The increased prevalence of robots in everyday life is threatening to human safety
Widespread adoption of robots in everyday life troubles me because it is blurring the boundaries between what is human and what is machine
Robots that appear life like are unsettling because they are almost indistinguishable from human beings
Recent advances in technology are challenging the very essence of what it means to be human
Technological advancements in the area of robotics are threatening to human uniqueness
Robots are beginning to blur the boundaries between what is human with what is machine

Feelings of power

This scale was adopted from Galinsky et al. (2003) and used in Experiment V, Chapter 6.

Please indicate how you experienced the task. To what extent were you...
...in charge of directing the task
...evaluating the robot's performance
...free to allocate power to the robot
...in a position of power over the robot

Appendix B

Other publications

C Bartneck, T Belpaeme, F Eyssel, T Kanda, M Keijsers, & S Sabanovic (2020). *Human-Robot Interaction. An Introduction*. Cambridge University Press

D Barry, M Shah, M Keijsers, H Khan, B Hopman (2019) xYOLO: A Model For Real-Time Object Detection In Humanoid Soccer On Low-End Hardware. In *Image and Vision Computing New Zealand conference Proceedings*

M Keijsers, C Bartneck, HS Kazmi (2019) Cloud-Based Sentiment Analysis for Interactive Agents. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (p. 43-50). New York, USA. doi: 10.1145/3349537.3351883

Ralston, C & Keijsers, M (2018) AI_r_t_school advertisement. In: *HAMSTER (2)*, ISSN 2538-0087

Ralston, C & Keijsers, M (2018) AI_r_t_school prospectus. In: *HAMSTER (3)*, ISSN 2538-0087

AI_r_t_ School welcomes queries from prospective students. In development are pioneering and innovative courses operating under the algorithms of creativity. Through random walks in databases of timeless art practices we rewire neural networks to the artistic discipline of your choosing. Unique human feedback unlocks pathways towards refined individual visual output.

___ART
___ART
___ART
___ART
___ART
___ART

___BOT
___BOT
___BOT
___BOT
___BOT
___BOT

> AI_r_t_ school

> robot learning

> range of courses available

> list__teach_

Modern
New Zealand
Landscape
Figure
Sculpture
Surreal
Performance

> AI_r_t_ school

> location_list

Christchurch
New Zealand

> human_input / critique_weekly

> course_time <6 week intensive>

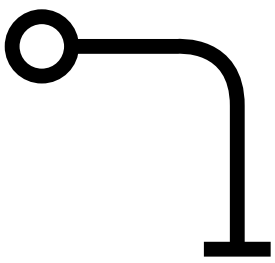
> prospectus available in
>> Issue Three of Hamster

> AI_r_t_ school

> 3.4GHz quad-core +

> language_supported_list

Lisp
Prolog
Python
C++
Perl



school

> contact
>> school.AI.r.t@gmail.com

Appendix C

Examiner feedback

C.1 Dr Alan Wagner

Research Weaknesses

1. Perhaps the most important concern is whether or not this work actually addresses the phenomenon of robot bullying and if this phenomenon actually exists. At the very end of the dissertation the researcher provides a rough definition of bullying:

   > *...bullying, there are also a few characteristics that emerge in the majority of them. Most scholars (see for example Ang & Goh, 2010; Casper et al., 2015; Gullone & Robertson, 2008; Hamburger et al., 2011; Jollie & Farrington, 2011; Modecki et al., 2014; Postigo et al., 2013; Sokol, Bussey, & Rapee, 2016) include the following components:*
   >
   > - *physical and/or psychological aggression that's intended and repeated over time;*
   > - *and which occurs in a dominant/submissive relationship, with a power imbalance between the bully (dominant) and the victim (submissive);*
   > - *and has the goal of harming or hurting the victim.*

   First, this definition should be stated at the beginning of the dissertation in order to establish the boundaries of what the work will encompass. Some of the characteristics of this definition are clearly present in the experiments. For example, 1) physical or psychological aggression and 3) with the goal of harming or hurting the victim both appear to be present in all experiments. You attempt to include the second characteristic, "which occurs in a dominant/submissive relationship, with a power imbalance" in experiments IV and V, but the manipulations seem to fail. One element of the definition seems to be missing for all experiments, "physical and/or psychological aggression that's intended and repeated over time." The fact that bullying occurs over time seems to me to be a critical component for bullying to occur. Moreover, the bullied and bully have a relationship (they know each other), interact repeatedly, and that this repeated interaction is difficult or impossible for the bullied person to escape are important elements that influence the learned helplessness that

results. Research shows that prestige and popularity motivate bullying behaviors in adolescents (Berger, C., & Caravita, S. C. (2016). Why do early adolescents bully? Exploring the influence of prestige norms on social and psychological motives to bully. Journal of Adolescence, 46, 45-56.). If we consider animal groups, dominance hierarchies are common and behaviors very similar to bullying are common (Kolbert, J. B., & Crothers, L. M. (2003). Bullying and evolutionary psychology: The dominance hierarchy among students and implications for school personnel. Journal of School Violence, 2(3), 73-91).

On the other hand, if we leave the two missing components, "repeated over time" and "power imbalance" what results seems to be little more than harmful, naked aggression. In fact, harmful naked aggression is commonplace. Consider, for example, the so-called "Knock-out game" in which one individual will sucker punch another individual trying to knock him out. This is certainly harmful naked aggression and does met criteria 1 and 3 above but does not include repeated interactions or a power imbalance. Generally, the knockout game is not considered bullying for this reason.

Questions

- Given the discussion above, should the phenomenon studied in this dissertation still be called 'bullying'? If so, how can the use of the term be justified given the discussion above? If not, then what is being studied and how do we motivate the topic?

[Candidate response] *I do believe the phenomenon described in this thesis should be called "bullying". As also recognised in the literature (Lowry et al., 2016; Volk et al., 2017) the power dynamic is a component of bullying, but not a prerequisite. If two people of equal standing meet, a power imbalance can be induced by either of those through aggressive behaviour towards the other, as long as the other gives in to this pressure. Of course, being in a pre-existing superior position will facilitate establishing a bully-victim relationship.*

Furthermore, one could argue that human-robot relationships tend to have a power imbalance implied: on one hand there is the human creator, on the other hand there is the mechanical minion, designed to carry out whatever tasks the human commands it to do. In Experiment I this power dynamic was shown through the auditor-auditee relationship. In Experiment II, the power imbalance existed because the participant could choose to put the robot in a humiliating situation through reviewing, and there would be nothing the robot could do about it.

Considering the repetition: I do believe this is indeed an important characteristic as it marks the difference between testing out the robot's capacities and limitations, and purposefully hurting it. However, as now also added to the definition in the thesis, repetition was incorporated in all experiments. In Experiment I participants saw fourteen videos of abusive behaviour towards the agent. In Experiment II.a the aggressor chose to continue exerting their power over the robot in spite of the protests; in II.b the questions asking about the acceptability of negative behaviour

repeatedly and for an extended period of time. In Study III any number of offences over 1 would imply repeated aggression. I re-ran the analysis with all the "number of offences" reduced by one with a minimum of zero, and found the same results. I will amend this in the corresponding chapter.

Regarding Experiments IV and V; in spite of all measures being merged into one single measure of aggression (i.e. the proportion of negative to positive responses), there was repetition in the abusive behaviour.

Then still, the difference between the 'knock out game' and my experiments is that individuals in the knock out game have consented to the violence. The robots in all my studies did not.

I have clarified the definition accordingly and moved it to the introduction.

2. In section 4.1.1 you discuss the ethical implications of chatbot abuse. You state, "accepting inappropriate and offensive language towards conversational agents could encourage users to abuse human interactive partners..." and you cite Brahnam 2005. I read this paper, curious about this statement. She does not make this claim in my opinion. The closest argument she makes in this regard is that chatbot abuse might bleed over towards product recognition. You state, "if we assume that users shape their behaviour to a bot based on their experience with fellow human interaction partners, it seems plausible that humans orient their behaviour towards other humans based on what behaviour is deemed acceptable towards robots and other nonhuman entities." This seems like a fallacy, even if we assume the antecedent, which is doubtable, the consequent does not follow. Yet the most dubious connection is the following, "Strait et al. (2018) indeed found that verbal aggression towards robots correlated with overall aggressive tweeting behaviour on Twitter, suggesting that abuse of robots and abuse of humans are related." Correlation of verbal aggression towards robots and aggressive tweeting behavior, however construed, does not in any way suggest that abuse of robots and abuse of humans are related.

   This entire section is an extraordinary and, at times, extreme stretch in scientific logic. There is little or no evidence at all that abuse of chatbots relates to abuse of anything else. If you look more broadly, you will find that there is actually significant evidence to the contrary. For example, there is a significant body of research looking into whether playing violent games leads to violent behavior. The argument has long been made, and closely relates to the argument you are trying to make, that abusing people in a game could possibly lead to abuse of real humans. Endless studies have argued for and against the hypothesis that playing violent games leads to violent behavior. For example, Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. Psychological science, 12(5), 353-359. But significant works also suggests that there is no relation: Ferguson, C.J. The Good, The Bad and the Ugly: A Meta-analytic Review of Positive and Negative Effects of Violent Video Games. Psychiatr

Q 78, 309–316 (2007). https://doi.org/10.1007/s11126-007-9056-9

I strongly suggest not entering this debate so haphazardly. The ethical reasons for preventing chatbot abuse are mostly illusionary and rely heavily on the opinions of a single author (Brahnam). Why not step into this work by simply motivating this research by stating that you are interested in understanding or reducing human frustration with chatbots? In other words, it does not seem necessary to motivate chatbot abuse prevention as an ethical issue.

Questions

- Why use ethical implications as a motive for this portion of your research? Do you believe and can you defend that it is unethical for a person to abuse a chatbot? If you claim that it is unethical to abuse a chatbot, then what about abusing your own computer, or verbally and/or physically abusing a stress squeeze ball? If the ball looks like a person does that influence the abuser to abuse people? Perhaps you can see how this line of argumentation can degenerate and become almost indefensible.

[Candidate response] *Human-chatbot interaction seemed like the right time to emphasize the ethical aspects of robot abuse, since the more pragmatic arguments (the robot might get damaged, which would be expensive to repair and may cause hazardous situations, and moreover keeps it from performing the task it was set out to do) do not hold up.*

I see the problem with the video game parallel but do not believe those two situations are comparable. When gaming, any violence or aggression is a means to an end; research (Przybylski et al., 2010) has shown that it's not the aggression in and of itself that makes games appealing or enjoyable but rather whether the challenges posed in the game give the player a sense of autonomy, competence, and relatedness. Even when violence is heavily integrated in the game, the goal always is to gain more points, collect more loot, progress another level; aggression is not a goal in and of itself. Relating back to the definition of bullying: the goal is not to hurt the victim and establish a power imbalance. This is a major difference with (persistent) abuse of chatbots, where the goal is to harm or hurt.

That being said, I do agree that this distinction needed to be made more clearly, and I have amended the paragraph to reflect this difference.

In addition, I would argue there is a major difference between a squeeze ball and an A.I. agent in terms of perception. Humans already perceive computers as social agents (the CASA framework, Nass et al., 1994; Reeves & Nass, 1996) but this gets even "worse" for robots (e.g. brain responses to a human versus robotic interaction partner are a lot more similar than when a human is compared to a computer interaction partner; Krach et al., 2008). Where exactly non-embodied agents are on this scale, I cannot say with certainty (I assume this will depend majorly on how convincing the agent is in mirroring human behaviour as well as the human's personal tendency

to anthropomorphise), but they will be considered social agents to some extent. This means that the ethical discussion is not about whether it is morally wrong for humans to abuse an object. The question is whether it is wrong for humans to abuse a social agent that cannot feel.

I cannot give an answer to that question. To me, it is similar to the ethical discussion about victimless crimes: we may be offended by the action even if no-one is directly harmed by it. This is also reflected in the public outcry about the inappropriate responses of (female) personal assistant A.I.s to sexual comments: the A.I. is recognised as representing a female social agent, and as a woman I can agree that it is infuriating to have such an agent respond in a meek or coy way to the type of abuse my own gender is battling so hard to have recognised as inappropriate.

The ethical debate on what behaviours we consider appropriate when it comes to non-living social agents will no doubt continue to be held. But this controversy in and of itself may already indicate that abuse of chatbots will be an important phenomenon to study, regardless of what side of the argument you are on.

3. Several of your experiments are conducted as virtual experiments with online human subjects. You follow some of these experiments up with in laboratory studies. Even so, experiment V suggests that embodiment is a factor that influences human aggressive behavior directed at a robot.

   Questions:

   - How should we evaluate the contribution of your online studies in light of results indicating that embodiment influences aggressive behavior directed at the robot? In other words, do the experiments conducted online only tell us about online bullying behavior? Do you believe that the results from these studies translate to physical situations involving an embodied robot? You mention some of the challenges associated with investigating robot bullying. What is the best way to investigate robot bullying?

   ***[Candidate response]** I added a section to the final chapter to discuss this a bit more in depth, but will try to give a short summary here.*

   I think that the incidence of aggression will differ between different embodiments as well as robot forms, due to factors such as self-awareness of the aggressor (very high in a controlled experiment in the lab, reasonably low in an anonymous online setting), robot cues (displays of emotion, coherency of social cues), et cetera. However, in line with research comparing on- to offline bullying, I do not think that the motivation for, or the fundamental nature of, robot bullying will be that different. Moreover, robot embodiments are and will be ever-evolving and changing. It thus makes no sense to focus on a single type of embodiment when studying robot bullying in the broad sense of the behaviour. Rather, by comparing results that were obtained from experiments with different kinds of robots and agents, we can try to distil stable findings out of

the overall variability. Those should be the results of interest, because they tell us something fundamental about HRI that will hold true across forms of embodiment.

Amendments

I suggest the following amendments:

1. The question of whether the candidate actually captures the phenomenon of bullying should be addressed early in the dissertation. The candidate will need to convincingly argue why the temporal component of the phenomenon can be ignored. Or alternatively change the subject of the dissertation to simply aggressive behavior directed towards robots.

2. The argument that it is unethical for a person to abuse a robot should either be significantly refined and developed, preferably in consultation with an ethicist or should be removed.

3. A small section, 2-3 pages, which reflects on which results translated from online studies to laboratory studies and which did not and why should be included in the conclusions.

4. The dissertation should be closely proofread to eliminate as many typos and grammatical mistakes as possible.

[Candidate response] *The amendments have been made; see also the responses to the individual points above. (Although the translation section as suggested in point 3 is about a page, rather than 2-3.)*

C.2 Dr Bilge Mutlu

Questions for Oral Defence

1. Overall, the dissertation presents "robot bullying" as phenomenon that has been observed in human-robot interaction and frames the research as inquiry toward better understanding the underlying mechanisms and factors of this phenomenon. Does the dissertation have secondary goals of informing robot design, the development societal guidelines (e.g., etiquette for conduct), or policy for legal or institutional response to these phenomena? If so, what are the specific implications of the dissertation for these potential outcomes? It is also appropriate for the dissertation to be a scientific inquiry toward better understanding the phenomena without any discussion of implications, but this goal needs to be stated explicitly, as there is ambiguity (e.g., Study III discusses this phenomenon affecting company- customer interactions). E.g., on Page 109 after reading Section 7.1, the reader may ask, "so what?"

   [Candidate response] *Amendments have been made to incorporate the long-term goal of helping to develop guidelines for robot design in both the introduction and the discussion. However, considering how much is unknown still, no solid recommendations can be made.*

2. Many of the studies are conducted using videos, simulated agents, and text-based descriptions; and involve third-party observations and/or anonymous interactions. While for each study appropriate controls have been created to ensure a reasonable level of realism and validity, at a high level it is unclear how much of what we learn from the studies will be applicable to the situations that described in Section 1 (Page 8) that motivated the research. What are appropriate steps toward bringing the research full circle?
   [Candidate response] *This point is very similar to the one raised by Dr Wegner (point 3, see above), so I will refer back to my response there as well as section 7.5.*

Additional Suggestions for Revision

1. There is a need to improve the consistency, flow, and accessibility of the dissertation document. It currently follows a "stapled together" dissertation format, where manuscripts that describe the five studies are brought together with common introduction and conclusion chapters. Although this is an appropriate format, and the dissertation does a reasonable job at following it, additional work is usually needed to connect the different studies together and to tell a coherent story. Specifically, the research questions provided in Section 1.1 (Page 9) and the hypotheses posited in Section 1.2.2 (Page 17) need to be connected with the research questions and hypotheses provided in each chapter. Additionally, some chapters provide explicitly stated research questions and hypotheses while others do not, and more consistency across studies can be established.
   [Candidate response] *Amendments have been made so that each chapter has an explicit set of research questions, hypotheses, and theoretical backup for the hypotheses.*

2. The Conclusion Chapter (Pages 4–122) can be extended in the following three ways:

   (a) By including a visual summary of how the research questions, central hypotheses, and findings from all studies are connected;

   (b) By discussing concrete limitations surrounding some of the questions asked above and the issues highlighted below (although "challenges" are discussed, it is difficult for the reader to understand the extent to which the findings can be applicable to real-world situations);

   (c) By adding a discussion of the design and social implications of the presented research.

   [Candidate response] *(a) I have added a table that provides a summary of the findings; a diagram would have gotten too complicated with the different findings and study designs. I hope the table provides an easy overview as well.*

   (b) These limitations have been merged with the Problems encountered with the concept "robot bullying", and the Implications sections (7.2 and 7.5, respectively).

   (c) A "Generalisations and implications" section has been added.

Comments & Questions on Individual Studies

In the following sections, I will list a number of lower-level and generally minor comments and questions for each study in order of appearance in the document (not in order of importance).

- *Study I*

  1. Page 25. The statement below needs further explanation. Why is variability across behaviors is a confound if the measurements are collapsed?

     Considering how the abusive behaviours covered a wide range of bullying behaviours, the possibility exists that one or more specific abuses would be considered unacceptable for one agent, but not the other. This would be a confound, as the fourteen measures are later on collapsed on a single index of abusive behaviour.

     [Candidate response] *The statement has been adjusted. Variability itself is not a confound; however, agent-dependent variability would have been. This would have indicated that some, but not all, forms of aggression are perceived differently depending on which type of agent is victimized. That would automatically reject hypothesis 1, "there is no fundamental difference in how people view robot versus human abuse"; but this effect might get lost if all data is aggregated without checking for interaction effects. Hence the confound testing.*

  2. Page 27. The formulation of "aggression" as an independent variable needs further explanation, as both the "aggressor" (human vs. robot) and whether it is provoked or not change across these groups. A more straightforward analysis would be conducting two separate models, one for human-to-robot aggression and one for robot-to-human aggression.

     [Candidate response] *I think there might be some confusion here, as the paragraph opens by stating that we did in fact run two separate models (or, more specifically, t-tests). The paragraph then continues to discuss the option of an ANOVA, and provides arguments why this would be an inappropriate test. I adjusted the text so that this distinction is more obvious.*

  3. Page 27. It is not clear why data for participants who found the material to be unrealistic was removed, as this can cause sampling issues and the removal of other important factors (e.g., people who find the videos unrealistic might also have particular opinions about robot bullying due to prior exposure, for example, to video games, etc.).

     [Candidate response] *Participants who though the material was unrealistic were removed because this arguably could bias the results. The aim of the study was to measure a response to robot bullying; if one does not believe that any bullying took place, then their response does not concern the acceptability of bullying.*

     However, to take away any remaining concerns, I would like to point out that, as stated in the "Exclusion of participants" paragraph, we actually also ran the

tests on both the whole dataset (i.e. with the participants who thought the videos unrealistic included) and a dataset where these participants had been removed. If the findings diverged, that is, if a significant effect became insignificant or vice versa, we reported both the results on the full dataset as well. This once, and is reported under "Acceptability of aggression towards agent" in the Main analyses.

4. Page 29. The first hypothesis must use a test of equivalence (or non-inferiority depending on what the exact hypothesis is). Please search for literature on equivalence and non-inferiority testing for appropriate methods (e.g., TOST Equivalence Test).
   [Candidate response] *With thanks for pointing this test out. I added the TOST equivalence test to the analyses.*

5. Page 31. The statement below needs clarification. Why would a human showing humanlike behavior would introduce a bias if the goal of the manipulation is to create "human" and "robot" agents?
   If the human actor had moved in a different way than the Atlas robot then this would have introduced another possible bias.
   Additionally, one possible explanation of the lack of differences in the acceptability of abusive behavior towards robots and humans, according to the earlier formulation in the chapter, is due to the reduced Human Nature cues in the human due to the robotic behavior. With reduced Human Nature cues, participants rated the acceptability at the (lower) level of a robot.
   [Candidate response] *A human showing humanlike behaviour might introduce bias through adding nonverbal cues that provide information about the victim's mental state, that would be absent with the robot. For example, if the human had slouched down following abuse, that would have provided the viewer with information about how the abuse affected them, which in turn might have altered their opinion on how acceptable it was. Thus, behaviour (nonverbal and otherwise) had to be identical between the agents in order to ensure that participants formed on opinion based on the same information.*

   *Additionally, the rigid movements of the human may have reduced how capable they were seen of experiencing feelings, but there were still major differences between HN traits attributed to the human (*M*(SD) = .80(.24), on a scale ranging from 0 - very incapable; to 1 - very capable) and to the robotic (*M*(SD) = .18(.21)) agent. It thus seems implausible that acceptability of abuse was due to both agents being perceived as low on HN traits.*

- *Study II*

1. Page 42. Is it possible that the "mind attribution" manipulation was not successful, and the introductions simply conveyed functional capabilities? I.e., saying that a robot has a mind might not increase mind attribution. Additionally, even if it increases mind attribution, it is possible that it changes other

aspects of the perception of the robot (e.g., its capabilities, its ability to mentalize, etc.).

[Candidate response] *The mind attribution manipulation introduction was tested in pilot 1, which used the same introduction as Experiment II.A. While I agree that it would have been better to include the manipulation check in the experiment itself, the pilot tested the same introduction and confirmed that the manipulation had been successful.*

However, since explicit mind attribution was measured in II.A, an extra manipulation check was carried out, which confirmed that people (explicitly) attributed more mind to the robot. I added the test to the Results section.

As can be seen in Table 3.3, the introduction took care to give the robot the same features: social behaviour, depth perception, object detection, etc. However, in one version this is framed as being the result of clever programming, while in the other version it is framed as the robot being sentient. This sentience of course will have come with qualities as mentalizing and perceiving the surroundings; but these are part of the mind attribution paradigm (see e.g. the questionnaire in Appendix A).

2. Page 44. Given the question above, would a mediation analysis fit better to the study design to account for the possible additional effects?

   [Candidate response] *It seems like a bit of an overkill to experimentally manipulate explicit mind attribution, then confirm this manipulation with a manipulation check (and later again in a pilot), and then furthermore assure that this manipulation indeed is the cause of the difference between the conditions through using the measurement that was used as a manipulation check as a mediator in a mediation analysis. But maybe three time's indeed the charm.*

   I thus ran three generalised linear models, to do a mediation. Model 1 is essentially the ANOVA in glm-format: dependent variable "condemning mistreatment" is predicted by independent factors "robot introduction" and "robot response". Both "explicit mind attribution in the robot introduction" and the "emotional response" increased how much participants condemned mistreatment. Model 2 had "mind attribution score" (as measured by the questionnaire) as dependent variable, and the same independent variables. Only "explicit mind attribution in the robot introduction" was positively related to the questionnaire scores, thus confirming that it can be used as a mediator. Model 3 was the same as Model 1, but added "mind attribution score" as continuous independent variable. The effect of the "robot introduction" factor disappeared, in favour of a significant effect of the mind attribution score (Sobel test: $Z = 2.41$, $p = .016$). This confirms that it was explicit mind attribution, and not some other third variable, that caused the increase of condemnation. (NB: the relationship between emotional response by the robot and condemnation remained significant).

3. Page 48. The statement below (and the associated action of excluding participants) need further justification and clarification.

If a participant failed to interpret Vector's response as negative, their response to the items assessing the unacceptability of that specific behaviour were not taken into account for the "unacceptability of robot bullying" score.
Data exclusion has to be carefully discussed and justified, as it can cause sampling issues as well as remove important effects that could otherwise be analysed. Could these people have low skill in social perception? If so, are the results only applicable with individuals with high skill in social perception?
[Candidate response] *So the awkward thing here is that I included those exclusion rules without actually checking if that meant that any data were excluded. This was not the case however – all participants interpreted shaking/lifting/verbally reprimanding the robot as negative. I amended the paragraph.*

4. Page 49. It is not clear why gender-balancing could not be established in the in-person study, as the researcher has control of participant recruitment (unlike in online studies). Additionally, given the large difference and the magnitude of the chi-square statistic (alpha level for co-variate exclusion is usually .25), including gender in the analysis as a covariate might be a better way to account for the potential effects of the sampling bias.
[Candidate response] *I am a bit confused. Gender balancing was established in the in-person study,* $\chi^2(2) = 3.96$, $p = .14$.

Regarding the option to add gender as a covariate in the online study. The point of randomisation is to make sure that any effect that side-variables like age, gender, etc. may have, directly or indirectly, on the outcome of interest (in this case, condemnation of robot abuse) will be similar between the conditions. If there is no relationship between any one of these variables and the outcome of interest, then this scenario has been averted, and including them as a covariate is not necessary.

What's more, including non-significant covariates might bias and obscure the main tests. These side-variables will "take up" degrees of freedom (thus reducing power) and can lead to overfitting. In addition, due to the missing data on age and gender for a proportion of the participants, including either of those as covariate would have greatly reduced power.

5. Page 50. A bit more discussion of the "public humiliation" behavior devised for Experiment II.B is needed. It is likely not seen as humiliating the robot by the participants, as writing or posting a negative review of a product, is not humiliation or abuse. This is similar with public reviews of people (e.g., ratemyprofessors.com).
[Candidate response] *This is an interesting suggestion and I have added the following points to the discussion of Experiment II as well. I think there is a difference between humiliating and bullying, and this is indeed one of the weaknesses of the experiment. In my opinion, it is very well possible to be condescending without intending to, just like it is very possible to make sexist or racist remarks (e.g. "you're pretty good at this for a woman" and "so, you're*

*living in [insert Western country] but where are you *really* from?") without intending to. This is also reflected in the pilot study: the condescending remarks were rated as significantly more demeaning. Even though participants may not have selected their review based on how much they wanted to humiliate the robot, that doesn't make the reviews less humiliating.*

However, it does pose a problem for the bullying definition, specifically the intentional part (this in addition to the repetition part). Thus, the bullying measure in Experiment II is by far the weakest of the measures in my thesis.

6. Page 51. The statement below requires a bit more discussion. An explanation of this finding is that one would expect high Human Nature to result in both the unacceptability of bullying and belittling in public, as one would not belittle an entity with low Human Nature (e.g., protecting the "little guy," the entity not understanding/appreciating belittling). One might belittle an entity that is worthy/aware of belittling; e.g., belittling a car achieves very little for the abuser.

   Exploratory analyses showed a positive relationship between finding robot bullying unacceptable and belittling it in public. This is intriguing, as common sense would suggest this relationship to be inverted.

   [Candidate response] *This is an interesting suggestion but in its current format it does not explain why the mind attribution manipulated acceptability of bullying but not belittling. Acceptability of bullying has to be dependent on more than just HN traits in order for this argument to work. However, this is very well possible and actually at the base of dehumanisation theory: it's not only HN that determines moral standing (as opposed to the Mind Attribution theory by Gray, who linked Experience/HN specifically to the right to be protected). I amended the discussion to include a possibility that the mind attribution manipulation acceptability through a combination of HN and – to a lesser extent – UH, while only HN was related to belittling behaviour.*

- *Study III*

1. Page 59. The dissertation indicates that "Study III harvested and content-analysed 283 conversations." What is meant by "content analysis?" Is it referring to the qualitative data analysis method (which the study does not follow) or only the data coding procedure?

   [Candidate response] *What was meant with "content analysis" was that the actual content of the conversations (rather than meta data like conversation length or the time it took people to formulate a response) was analysed. However, this phrase was confusing and has been taken out.*

2. Page 60. Was this study conducted under IRB supervision or did the research rely on the terms & conditions of the Cleverbot website?

   [Candidate response] *The study was relying on the terms & conditions of the Cleverbot website, the existence of precedents (De Angeli & Brahnam, 2008;*

De Angeli & Carpenter, 2005; Hill et al., 2015), and the fact that no information other than the conversation itself (e.g. IP address, date or time of the conversation, time zone, duration, etc.) was collected.

3. Page 66. Does the statement "humanlikeness of the chatbot on the other" refer to "claims of humanity" by the chatbot that has been coded in the data?
   [Candidate response] *Partially. As specified in 4.1.2, humanlikeness of the chatbot was measured through more than one aspect. In addition to the claims of humanity, there were the approximate Loebner score (i.e. how many naive readers would mistake Cleverbot for a human) and number of nonsensical responses by Cleverbot (negatively related to humanlikeness).*

   Also in response to the reviews that the paper which this chapter is based on received a month ago, this chapter has been revised, and hopefully gotten clearer in its methods and interpretation.
4. Page 66. The statement "it was hypothesised that" the chapter states hypotheses earlier, but no hypotheses are provided.
   [Candidate response] *Hypotheses have been provided.*
5. Page 69. An important limitation of the study is that the interactions with the chatbot are anonymous. To what extent will the results from this study generalize to interactions where the identities of the users are known to a third party (e.g., the chatbot company, the public)? Much of the studied antisocial behavior will likely significantly diminish, and the quantified relationships might change. It is important to outline the limitations on the applicability of the findings to user-chatbot interactions.
   [Candidate response] *This is discussed to some extent in the virtual studies introduction (5.1.1). I expect that the anonymous online setting may have resulted in higher incidence (although a non-anonymous human-chatbot interaction study found similar rates of sexual references, see Table 4.3). However, just like cyberbullying and in-person bulling do not differ in principle, I do not think that there is a fundamental difference in motivation for the abuse of chatbots when people are anonymous versus when their identity can be uncovered by a third party.*

   I have included my argument to the limitation section in the discussion of Chapter 4.
6. Page 70. Additionally, it would be important to discuss how these findings might apply to human-robot interactions, as this is the main topic of the dissertation.
   [Candidate response] *Mostly see above. I have amended this in the limitation section.*

- *Study IV*

1. Page 74. The humanlikeness manipulation as described in the text below may not only manipulate humanlikeness but may also manipulate congruency be-

tween the robot's appearance and behavior. More discussion on the potential externalities should be included.

The robot either had a humanlike voice and gave off social cues through movement (high humanlikeness condition) or spoke with a synthesised voice and was shown in stills (low humanlikeness condition).

[Candidate response] *I am not sure if there would be that much incongruency to be honest. The robot has both humanlike features (humanoid shape, face) and robotic features (colour, cartoonised features). Both the humanlike and the computer-generated voice fitted well, I think. But this was not pilot tested. Humanlikeness may have been an ill-chosen name for the conditions though, as I would rather call it "presence of social cues" now. I have added this point to the limitations section.*

2. Page 75. Although the use of the interactive, non-linear stories is a strength in terms of offering more realistic and varied interactions, it also can create variability that can overshadow studied effects. One way to account for this variability is to model the robot's behaviors as covariates. See Peltason et al. 2012 for an example of how this can be done.

   [Candidate response] *As stated at the end of the Reliability, randomisation, and manipulation check (section 5.2.2, Results), on average 75% of the participants' interaction paths overlapped. While there will be some error introduced through the variability, I don't think this is too problematic with such an overlap. Moreover, adding the behaviours as covariates would be not as simple as that. The "decision tree" of interactions looped and linked between branches: there were 70 passages, connected through 139 links. A participant could take more than one path to end up at the same interaction, which would likely affect the influence of that interaction. Moreover, adding all these 70 passages as covariates would create an impossibly over-fit model.*

   In short, I appreciate the suggestion, but will not adjust for the variability.

3. Page 78. The statement below suggests a causal relationship between two outcome variables. Whether a causal relationship is being studied should be clarified.

   The current experiment set out to empirically test whether reduced mind attribution to (i.e. dehumanisation of) a robot causes a greater proclivity of people to bully the robot.

   [Candidate response] *The paragraph has been amended and now includes a sentence on how due to the manipulation failure only correlational relationships can be inferred.*

4. Page 79. There is a need to theoretically ground the power manipulation, as asking participants to imagine themselves as being president was not successful and resulted in unexpected behaviors such as criticizing particular political figures. In the management literature, power is usually associated with relative organizational roles and impact on the future benefits to the individual. Al-

though the power manipulation in latter experiments are more closely aligned with this formulation, there is a need to better ground and discuss this factor. ***[Candidate response]*** *To be quite frank, I don't see what amendments are suggested here. The power manipulation is theoretically grounded in the dehumanisation literature, as explained in 5.1.1 and further discussed in 5.2.3. Clearly, the initial attempt at manipulating feelings of power was too liberal an interpretation of the original (which was taken from the literature). This is recognised in the Discussion section of study IV.a (section 5.2.3) and then amended for study IV.b, as described in 5.3.1 under "procedure, materials, and measurements". Then finally, the overall discussion section once more discusses the failed manipulation and its grounding in previous studies as a limitation. I don't know how to further stress that the measure was theoretically grounded, and based on previous studies reported in the literature, without getting repetitive.*

- *Study V*

  1. Page 88. The hypotheses would be more accessible if they are presented in a more formal fashion (enumerated, provided with justification).
     [Candidate response] *Hypothesis presentation has been amended.*
  2. Page 90. The embodiment manipulation involves manipulating many more factors than just agent embodiment that I do not think that this formulation of the study and the discussion of its results is valid. This is a major limitation of the study the requires addressing by reformulating the studies. One possibility is to analyse the data from the online and in-person studies separately and make high-level comparisons between the effects rather than direct statistical comparisons between the datasets. For example, Page 100 states "More or less in line with expectations, people were kinder to an embodied robot than to a virtual one." The in-person nature of the study likely has a stronger effect on this outcome than the physical embodiment of the robot. ***[Candidate response]*** *The problem with splitting up the experiment in two separate analyses is that a lot of power would be lost; this would be acceptable for the virtual experiment at .85 but rather low for the embodied robot at .56. Through combining the two and adding embodiment as a factor, the power to detect an effect increases; and as interaction effects are included, the possibility that the effect found in the virtual condition wrongfully "carries over to" the embodied condition is eliminated.*

     Of course, this leaved the issue of interpretation. I agree that there are differences between the virtual and embodied robot condition beyond the mere embodiment of the robot, and maybe as a result the label shouldn't be virtual vs embodied, but rather cyber- vs in-person bullying. I have added a discussion of this in the limitations section.
  3. Page 103. The section on "future work" should be removed or worked into the

discussion, as it is not appropriate for a dissertation.
[Candidate response] *The "future work" section has been merged with the Future Research section in the Conclusion.*